informs
https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–16
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

How the Design of Ranking Systems and Ability Affect Physician Effort

Katharina Huesmann,[a] **Yero Samuel Ndiaye,**[b,c] **Christian Waibel,**[d] **Daniel Wiesen**[e,f,*]

[a] School of Business and Economics, University of Münster, 48143 Münster, Germany; [b] Department of Economics, University of Cologne, 50923 Cologne, Germany; [c] Max Planck Institute for Research on Collective Goods, 53113 Bonn, Germany; [d] Independent Researcher; [e] Department of Operations Management, University of Cologne, 50923 Cologne, Germany; [f] Erasmus School of Health Policy and Management, Erasmus University Rotterdam, 3062 PA Rotterdam, Netherlands

*Corresponding author

Contact: katharina.huesmann@wiwi.uni-muenster.de, https://orcid.org/0000-0001-5210-2752 (KH); ndiaye@wiso.uni-koeln.de, https://orcid.org/0009-0002-9646-6923 (YSN); christian.waibel@gmail.com, https://orcid.org/0000-0002-0162-7644 (CW); wiesen@wiso.uni-koeln.de, https://orcid.org/0000-0003-4627-1730 (DW)

Received: April 6, 2022
Revised: February 29, 2024; October 28, 2024
Accepted: November 16, 2024
Published Online in Articles in Advance: May 30, 2025

https://doi.org/10.1287/mnsc.2022.00990

Abstract. Although relative performance feedback in the form of rankings appears to be effective in improving health outcomes, it may have either motivating or demotivating effects for individual physicians. Potential factors influencing such effects include a physician's ability and the design of the ranking system itself; however, there is limited understanding of these factors. Using a controlled lab-in-the-field experiment with practicing and future physicians as subjects ($N = 352$), we systematically analyze effort within small teams under different ranking systems. Exogenously varying the number and position of the thresholds defining the ranking system, we observe that the addition of a threshold to create a new rank is motivating—that is, increases effort—only among individuals capable of exceeding that threshold; the effort of other individuals may remain unchanged or even decrease. In particular, a highly granular ranking system with ranks spanning the entire range of possible outcomes maximizes overall physician effort; high thresholds serve to motivate high-ability individuals, whereas moderate and low thresholds provide opportunities for improvement to lower-ability individuals who cannot reach the high thresholds. Our results suggest that, to motivate their teams effectively, clinical leaders should provide rank feedback using a system under which physicians of all ability types can improve their rank through increased effort.

History: This paper was accepted by Stefan Scholtes, healthcare management.

Funding: This work was supported by C-SEB Center for Social and Economic Behavior at the University of Cologne (Junior Startup Grant No: Rd04-2016-JSUG-Huesmann) and the School of Business and Economics at the University of Münster (Junior Researcher Promotion Fund). This research was initiated while Daniel Wiesen was employed at the Institute of Health and Society at the Department of Health Management and Health Economics, University of Oslo (Norway), supported by the Research Council of Norway [IRECOHEX, Grant No. 231776]. Y. S. Ndiaye gratefully acknowledges support by the German Science Foundation through Germany's Excellence Strategy [EXC 2126/1 390838866].

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2022.00990.

Keywords: ability • lab-in-the-field experiment • rankings • relative performance feedback • status concerns

1. Introduction

Improving the quality of care is a key objective for hospitals. For clinical leaders, one important aspect of doing so is to motivate individual physicians to provide high-quality care. To this end, professional medical societies increasingly advocate the use of peer feedback in clinical settings—in particular, the practice of informing physicians about their performance relative to that of their peers (as measured, for example, by clinical indicators; see Valori et al. 2018, Siau et al. 2019). Feedback of this kind appears to be especially relevant in clinical fields with high volumes of activity and measurable indicators of individual performance, such as gastroenterology (Corley et al. 2014). The basic logic is that relative performance feedback increases the salience of social comparison, which drives individuals to intensify their efforts (see, e.g., Roels and Su 2014, Gill et al. 2019). However, some studies warn that peer feedback may give rise to heterogeneous and potentially negative responses (see, e.g., Bandiera et al. 2013, Charness et al. 2014, Turkoglu and Tucker 2022).

One standard way in which relative performance feedback may be expressed is in the form of rankings shared among clinical team members. A ranking system is a collection of *thresholds* dividing the range of possible health outcomes (for a given field or team of physicians) into *ranks*. From the perspective of a clinical leader, the design of a ranking system is a nontrivial task because different physicians may respond to rank feedback in different ways. Those who have a chance of reaching a higher rank may be motivated to work harder to do so, whereas those who have no chance of reaching it may be demotivated by their failure to do so. Because physicians vary in ability, this creates a potential tradeoff. The inclusion of a particular rank in a system may motivate some physicians on a team while demotivating others. (By "ability" here, we mean an individual's capability to perform a particular task or activity, which reflects their inherent talent, training, and experience. Variations in ability mean that different physicians investing the same level of effort in diagnosis and treatment may achieve different health outcomes; see, e.g., Chan et al. 2022, Gowrisankaran et al. 2023.) A clinical leader must therefore answer the following question: How many thresholds should the ranking system contain, and how should they be distributed within the range of possible outcomes?

As a concrete example, consider a ranking system based on adenoma detection rates, which are an essential quality indicator in gastroenterology.[1] Adenoma detection is a high-volume activity, and rates are measurable at the individual-physician level, making them a good candidate for a performance measure.[2] According to the United Kingdom's Joint Advisory Group on Gastrointestinal Endoscopy, for example, adenoma detection rates between 10% and 20% indicate adequate colonoscopy quality. In this context, if the threshold for the top rank in the system is very high, say 25%, this rank will be motivating for a few physicians who can reach it but potentially discouraging for the many who cannot. On the other hand, a low threshold of 10% will provide no additional motivation to the vast majority of physicians because they can meet it without expending extra effort.

In this paper, we analyze how the design of a ranking system affects physician effort levels and how these effects depend on individuals' abilities. Our study illuminates the mechanisms behind the heterogeneity in individuals' responses to relative performance feedback (for more details, see Section 2). We use a well-powered and preregistered controlled lab-in-the-field experiment, with 112 physicians working in inpatient care and 240 future physicians (medical students) as subjects. (A lab-in-the-field experiment follows a standardized laboratory paradigm but is conducted in a naturalistic setting; see, e.g., Gneezy and Imas 2017.) To the best of our knowledge, ours is the first controlled, incentivized experiment on relative performance feedback conducted with both practicing and future physicians. The experimental design is well grounded in theory because we base our behavioral predictions on an economic model of status concerns that incorporates status utility (Moldovanu et al. 2007; see Online Appendix A). Subjects make stylized decisions about a series of abstract healthcare tasks, with their effort choices reflecting tradeoffs between incurred costs and patient health benefits as well as status concerns induced by rank feedback. Before conducting the experiment, we interviewed seven clinical leaders from German hospitals, who validated the practical relevance of our study and confirmed that the stylized experimental design captures physicians' real-life incentives.

Our behavioral findings carry important implications for clinical leaders who are considering using rank feedback to improve performance in their teams. To motivate physicians optimally, across all ability levels, our results suggest that clinical leaders should opt for a ranking system with thresholds spanning the entire range of possible outcomes. A system with only a single threshold near the top of the range of outcomes, demarcating a single high rank that only high-ability physicians can reach, will demotivate lower-ability individuals. Nonetheless, the system should contain a threshold near the top to motivate high-ability physicians to increase their effort. However, to avoid demotivating low-ability physicians, the system should also contain lower thresholds all the way to the bottom of the outcome range. This gives all physicians the opportunity to improve their rank, irrespective of their level of ability.

Our experiment is structured as follows. First, all subjects make effort choices for an initial set of tasks before they have been tested on their ability or received any feedback. This establishes a baseline without feedback. They then take a test in order to assign them to an ability type (either *high* or *low*) in their group. In the focal part of our experiment, they make effort choices under each of five ranking systems. These choices translate (stochastically) into patient health outcomes, with the set of outcomes each individual can achieve being dependent on their ability type. At the end of the experiment, one randomly chosen ranking system is made public among peers in a group. This makes social comparison a salient factor in subjects' choices.

In our stylized setup, the set of possible outcomes has just four elements, so it can contain at most three thresholds: a *top* threshold (separating the highest outcome from the rest), a *middle* threshold (separating the top two outcomes from the bottom two), and a *bottom* threshold (separating the lowest outcome from the rest). Only high-ability subjects can meet the top threshold, whereas only low-ability subjects can fail to meet the bottom threshold; all subjects can meet the middle threshold. The five ranking systems we consider correspond to various combinations of these thresholds.

We find that the subjects' effort choices do indeed depend on the ranking system, and the relationship between effort and ranking system depends on the subject's ability type. High-ability subjects choose the highest levels of effort under the two ranking systems that include the top and middle thresholds. Similarly, low-ability subjects choose the highest levels of effort under the two ranking systems that include the thresholds they can reach, namely, the middle and bottom thresholds. (However, they are slightly demotivated by the inclusion of the top threshold.) In aggregate, the most granular ranking system—the one with all three thresholds—results in effort levels 5%−25% higher than those under the other systems. The system consisting of only the top threshold yields the lowest effort levels, 13%−20% less than the other ranking systems.

Compared with the baseline without feedback, we find that high-ability subjects expend significantly more effort when faced with a ranking system that includes the top threshold but not much more under a system that does not. In contrast, low-ability subjects never expend more effort, compared with the baseline, under any ranking system; effort decreases to a large extent under systems that do not include both the middle and the bottom threshold. In aggregate, the most granular ranking system induces significantly higher effort (about 5%) compared with the baseline, whereas the system consisting of only the top threshold induces about 13% lower effort.

2. Literature and Hypotheses

2.1. Related Literature on Relative Performance Feedback

This paper contributes to the literature in healthcare management and behavioral economics on the effects of relative performance feedback (in the absence of financial incentives). Some studies report that relative performance feedback has positive effects on performance in healthcare organizations. For example, Song et al. (2018) showed that public disclosure of relative performance information on the length of stay of high-acuity patients, along with sharing of best practices, increases productivity in emergency departments. Navathe et al. (2020) reported that relative performance feedback improves quality in primary-care organizations. Niewoehner and Staats (2022) found that performance feedback at the hospital level increases flu vaccination rates more than financial incentives do.[3]

A number of behavioral experiments also address the effects of relative performance feedback. Public feedback has been found to improve performance by giving rise to social comparison among ranked peers (see, e.g., Hannan et al. 2013, Tafkov 2013, Gerhards and Siemer 2016), whereas private feedback may do so by stimulating people's self-image concerns (see, e.g., Tafkov 2013, Gill et al. 2019). Kuhnen and Tymula (2012) observed an ex ante effect; when individuals learn in advance that rankings will be announced, they increase their effort.

Many studies, however, report that relative performance feedback has negative or null effects on performance (see, e.g., Bandiera et al. 2013, Ashraf et al. 2014, Charness et al. 2014, Edelman and Larkin 2015, Turkoglu and Tucker 2022); see Schnieder (2022) for a review of the experimental literature. Singh and Zureich (2024) showed that the performance of clinical physicians may improve in response to positive feedback but deteriorate in response to negative feedback.

In light of these rather mixed findings, it is important to understand better the sources of heterogeneity in individuals' responses to rank feedback. Surprisingly, the literature has not systematically considered the design of the ranking system as a potential source of heterogeneity. Most studies compare performance in a situation with no feedback to performance under one specific ranking system—typically either a fully granular system (with one rank per outcome) or a system that honors only top performers. An exception is the experiment of Hannan et al. (2008), which considers two types of ranking system: coarse (individuals are privately informed of their position relative to the median performance) and fine (individuals are informed about their performance percentile). Hannan et al. (2008) reported that private feedback improves performance, but they found no significant difference between the results under coarse and fine ranking systems. Also, their experiment simultaneously implements financial incentives, which makes it difficult to isolate the effect of the type of ranking system.

Response heterogeneity may also be related to individuals' previously achieved ranks. For example, many studies have documented the phenomena of first-place loving and last-place aversion, in which individuals who achieve very high or low ranks show particularly large effort increases afterward (Azmat and Iriberri 2010, Kuziemko et al. 2014, Gill et al. 2019, Niewoehner and Staats 2022). Correspondingly, Turkoglu and Tucker (2022) found that receiving feedback causes the performance of middle-ranked individuals to suffer. On the other hand, Bradler et al. (2016) reported on an experiment in which individuals who did not achieve the highest ranking drove most of the subsequent performance improvements. Similarly, individuals ranked last may become demotivated and prone to giving up (Müller and Schotter 2010, Buell 2021, Cotofan 2021). These studies generally considered repeated-decision situations under a single design of ranking system. This raises the question of whether individuals of various ability levels may react differently to feedback depending on the ranking system in play.

If rank feedback is provided through repeated decisions or in contests, it may also inform individuals about their chances of winning a prize (Dechenaux et al. 2015). This complicates the question of how feedback affects effort. In our study, we avoid this complication by providing feedback only at the end of the experiment so that the content of the feedback cannot affect subjects' decision-making within the experiment. That is, we focus on the ex ante effects of feedback (Kuhnen and Tymula 2012, Coffman and Klinowski 2025). In addition, because our subjects know their ability types, they are fully informed about their chances of achieving each possible outcome (and hence, each rank). This lets us clearly distinguish the effects of effort from those of ability, which is not usually possible in field settings (Ericsson and Charness 1994).

Our central contribution to the literature is our systematic analysis of how the design of a ranking system affects effort across different levels of ability. Unlike previous studies, we compare subjects' behavior under an essentially comprehensive set of ranking systems (for the stylized situation in our experiment). Our paper is the first to make such within-subject comparisons and to break down the interaction between individuals' responses to rank feedback and their levels of ability.

2.2. Hypothesis Development

We now formulate two hypotheses to test in our controlled experiment. Specifically, we are considering a form of feedback in which each physician on a team is assigned a rank based on his or her performance in a clinical activity, that is, based on the *outcome* of that activity. The map from outcomes to ranks (which is the same for all physicians) is called the *ranking system*; it is determined by *thresholds* placed within the set of possible outcomes, which demarcate the ranks. Physicians with the same outcome are assigned the same rank. For example, all physicians achieving outcomes above (below) the highest (lowest) threshold are assigned to the first (last) rank. The challenge for the clinical leader designing the ranking system is to decide where to set the thresholds. For insight into this challenge, we examine how physician effort is affected by *adding a threshold* to a given ranking system so that all physicians with outcomes above the new threshold retain their previous ranks, whereas those with outcomes below the new threshold are assigned a lower rank.

Our reasoning is based on the fact that, in the absence of financial incentives, rank feedback influences behavior by prompting social comparison (see, e.g., Suls and Wheeler 2000, Brown et al. 2007, Tafkov 2013, Gill et al. 2019). Higher ranks represent higher status within a team and so yield higher utility (Zizzo 2002). Adding a threshold to the ranking system thus gives individuals more opportunities to stand out, which may motivate them to increase their effort. Whether it actually motivates them, however, likely depends on their ability to meet the new threshold. The addition of a too high threshold might discourage individuals unable to reach it. Such an effect would be consistent with the observation from the tournament literature that the offer of a reward can reduce effort from individuals who are unlikely to win it (see, e.g., Hannan et al. 2008, Newman and Tafkov 2014). On the flip side, a threshold that is too low may fail to motivate high-ability individuals (e.g., highly experienced physicians) because they need not fear falling below it.

To make these arguments rigorous, we consider a model of status concerns that incorporates status utility (Moldovanu et al. 2007, Dubey and Geanakoplos 2010). The core assumption is that an individual's status utility depends positively (negatively) on the number of individuals ranked below (above) them. (For a formal description of the model, see Online Appendix A.) Therefore, adding a new threshold affects the status utility associated with outcomes near that threshold: It increases the status utility for outcomes just above the new threshold and decreases it for those just below. This means that a slightly lower-ranked individual who can reach the rank above the new threshold is likely to increase effort to try and do so because that rank has become more valuable and staying in a lower rank has become more painful. Conversely, individuals who cannot reach the new threshold are hurt by its addition because their outcomes are no longer pooled with the higher ones above the new threshold; such individuals are likely to decrease their effort. Furthermore, individuals who can easily surpass the new threshold (either because they have high ability or because the threshold is very low) are also likely to decrease their effort because they can now obtain the same utility as before, but with less effort. An individual's response to the addition of a new threshold thus depends on their ability to reach it. These observations, which are formalized in Proposition EC.1 in Online Appendix A, lead us to the following hypothesis.

Hypothesis 1 (Ranking System Design and Ability). *Adding a threshold to a ranking system will affect individuals' effort choices. The direction of the effect for a given individual depends on whether that individual can meet the new threshold.*

a. *For individuals who can reach outcomes both above and below the new threshold, effort increases.*

b. *For individuals who cannot reach outcomes above the new threshold, effort decreases.*

c. *For individuals who cannot reach outcomes below the new threshold, effort decreases.*

Because physicians within a team may vary in ability (e.g., because of differences in experience or training) and all face the same ranking system, Hypothesis 1 highlights the potential tradeoffs in designing a ranking system. Thresholds placed in the middle of the outcome range, which one might expect to be attainable (yet somewhat challenging) for all team members, should affect them all positively. However, thresholds near the top of the range, which may be attainable by only a few team members, may motivate those few but demotivate the rest. Similarly, thresholds near the bottom of the range, which may be trivial for most team members to meet, may motivate the few who may fear to fall below them but demotivate the rest. The *empirical* question is whether the positive effects of adding an extremely high or low threshold outweigh the negative effects.

According to the literature, individuals respond to rankings in a nonlinear way, and the prospect of being ranked either first or last in a group is particularly motivating (see, e.g., Azmat and Iriberri 2010, Müller and Schotter 2010, Newman and Tafkov 2014, Gill et al. 2019, Buell 2021, Niewoehner and Staats 2022). In addition, the scarcity of a reward makes it more attractive (Besley and Ghatak 2008). It is therefore reasonable to suppose that, in our setting, an extremely high threshold is strongly motivating (to those for whom it is within reach) precisely because few people can reach it. Likewise, in the presence of an extremely low threshold, individuals who might miss the threshold are strongly motivated to avoid doing so, precisely because almost everyone else will surpass it.

In other words, when the top and bottom ranks are defined by extreme thresholds, reaching the top rank and avoiding the bottom rank should become particularly attractive. Therefore, we expect that the motivating effects of these thresholds are higher than the demotivating effects. More specifically, we hypothesize that adding a high threshold motivates individuals who can reach it more than it demotivates those who cannot. Likewise, a low threshold motivates those who might fall below it more than it demotivates those who never will. When adding both a high and low threshold, we expect an unambiguous increase in effort (i.e., effort increases for those who can reach the top rank as well as for those who can reach the bottom rank).

Hypothesis 2 (Salience of extreme thresholds). *If a threshold is added near the top (bottom) of the outcome range, then the resulting increase in effort from individuals who can attain outcomes above (below) that threshold will be greater than the decrease in effort from individuals who cannot. Adding thresholds near both ends of the outcome range increases effort for both types of individuals.*

As described in Section 2.1, the literature has established that relative performance feedback can have both positive and negative effects. Thus, by testing Hypotheses 1 and 2, we advance the literature by disentangling these effects and explaining each one in terms of the design of the ranking system.

In addition, our experimental design allows us to compare effort choices in the presence of rank feedback to a baseline with no rankings. We can therefore check (in our setting) the effect of providing relative performance feedback compared with not providing feedback.

3. The Experiment

3.1. Recruitment and Power Analysis

We conducted our experiment between May 2023 and January 2024. The experiment obtained ethics clearance from the German Association for Experimental Economic Research (No. gzKUnEzB) and from the Ethics Committee of the Faculty of Management, Economics, and Social Sciences at the University of Cologne (No. 230015DW). It was preregistered on the platform AsPredicted (No. 130723).

In total, 112 physicians working in inpatient care and 240 medical students participated in the experiment. Recruitment of physicians was facilitated by hospital department heads, who sent emails to their clinical teams asking them to participate. The experiments with physicians were conducted in seven hospital departments, spanning five hospitals, in western and southern Germany. Medical students were recruited by email through the office of the Dean of the University of Cologne Faculty of Medicine, and the experiments with medical students took place there. For the experiment procedure, see Section 3.4.

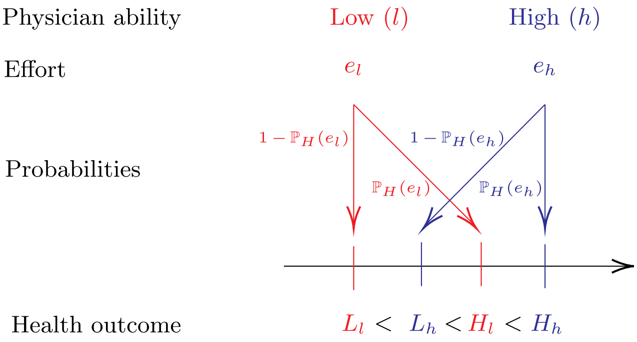
To determine the sample size for within-subject comparisons of ranking systems, we conducted an a priori power analysis. We assumed a medium effect size (Cohen's $d = 0.4$), a conventional power of 0.8, and a statistical significance level of $\alpha = 0.05$ (Cohen 1988). Using a Bonferroni correction and nonparametric two-sided Wilcoxon signed-rank tests, we obtained a sample size of 112 subjects per experimental treatment.[4]

3.2. General Design and Decision Situation

Our experiment is framed as a series of stylized healthcare decisions. In each task, the subject chooses an effort level $e \in \{0, 1, 2, \ldots, 10\}$, at cost $c(e) = 2e$, to treat an abstract patient. (Although the stated-effort paradigm (in which subjects simply state what effort level they will invest, rather than actually performing an activity requiring effort) has certain limitations (see, e.g., Charness et al. 2018), it nevertheless allows us to capture a physician's concern for patient health, profit, and status.) For each task performed, the subject receives a lump sum of 20 ECU, the experimental currency; this amount does not depend on the realized health outcome. The subject's profit is thus $\pi(e) = 20 - c(e)$. We provide exchange rates of 1 ECU = 3 euro for subjects who are working physicians and 1 ECU = 0.8 euro for subjects who are medical students.

In each task, the patient's health outcome depends on the subject's effort and ability type. Each subject's ability type is either high (h) or low (l). High-ability subjects will achieve either a high outcome H_h or a low outcome L_h; low-ability subjects will achieve either a high outcome H_l or a low outcome L_l. We assume that $L_l < L_h < H_l < H_h$. (For an illustration, see Figure 1.) For either type, the probability of a high outcome (H_h or H_l) is $\mathbb{P}_H(e)$, which is an increasing function of the individual effort e chosen. Accordingly, for either type, the probability of a low outcome (L_h or L_l) is $1 - \mathbb{P}_H(e)$. In our main experimental treatment (hereafter labeled as MAIN), we quantified the possible outcomes as $L_l = 0$, $H_l = 20$, $L_h = 5$, and $H_h = 25$; see Table B.1 in Online Appendix B for the experiment

Figure 1. (Color online) Ability Types, Effort, and Health Outcomes



Notes. This figure shows how subjects' effort choices translate into health outcomes, depending on their ability type. If a low-ability subject chooses effort e_l, they achieve outcome H_l with probability $\mathbb{P}_H(e_l)$ and outcome L_l with probability $1-\mathbb{P}_H(e_l)$. If a high-ability subject chooses effort e_h, they achieve outcome H_h with probability $\mathbb{P}_H(e_h)$ and outcome L_h with probability $1-\mathbb{P}_H(e_h)$.

parameters. To test the robustness of our parameterization, we also ran a Control treatment with $L_l = 10$, $H_l = 20$, $L_h = 15$, and $H_h = 25$.

Although the tasks in the experiment deal with abstract patients, we incorporated the factor of a physician's concern for real patients by translating the health benefits from each subject's decision into monetary terms and transferring this amount to the Christoffel Blindenmission, a charitable organization, to be used for the treatment of cataract patients.[5]

The experiment was administered in computerized form, using the platform oTree (Chen et al. 2016). Subjects were randomly assigned to groups of four, which remained constant throughout the experiment. Each group was seated at a table with four laptop computers.

The experiment took place in three stages. In the first stage, subjects completed a task in the absence of a ranking system and without knowing their ability types. Following a strategy-method format (Selten 1965), each subject made two effort choices: one assuming that their ability type was high and one assuming that it was low. This gave us a baseline for each subject's effort before the introduction of rankings.

In the second stage, we determined each subject's ability type by asking them to answer nine questions from the German admissions test for medical studies (Test für Medizinische Studiengänge); see Online Appendix B. In each group of four subjects, the two with the fewest correct answers were identified as low-ability and the other two as high-ability (with ties broken at random). We then privately informed each subject of his or her ability type and of the outcomes that he or she would thus be able to achieve. (We assigned ability types using a real-effort task, rather than an arbitrary method, in order to stimulate status concern.)

After the test, subjects were asked to introduce themselves within their group of four by calling out their first names, then to type their names into their computers (so that their rankings could be displayed at the end of the experiment). This procedure makes each subject's identity public within the group (Rege and Telle 2004, Loch and Wu 2008).

In the third stage, subjects made effort choices under five different ranking systems (described in Section 3.3). We used a one-shot decision setup, rather than repeated decisions, in order to focus on the ex ante effects of the prospect of rank feedback (as opposed to the effects of feedback content on future decisions). The five ranking systems appeared in a random order on each subject's screen. After all subjects had made their effort choices, one of the five ranking systems was randomly implemented for each group; subjects' ranks were then publicly disclosed among the individuals in the group.

3.3. Ranking System Designs

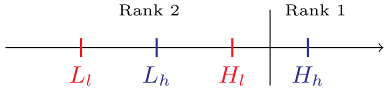
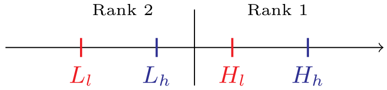
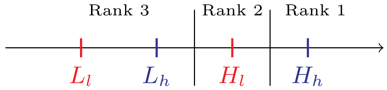
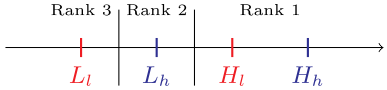
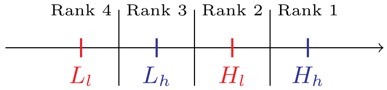
As described in Section 2.2, a ranking system is a map from the set of all possible outcomes to a set of ranks; it is determined by a collection of thresholds placed within the set of outcomes, which demarcate the ranks. In our experiment, a task has four possible outcomes, $L_l < L_h < H_l < H_h$, so there is room for up to three thresholds. For convenience, we give names to these potential thresholds; the *top threshold* lies between H_h and H_l, the *middle threshold* between H_l and L_h, and the *bottom threshold* between L_h and L_l.

Table 1 depicts the five ranking systems that we test in our experiment. Under the *T ranking system*, given by the top threshold alone, subjects achieving outcome H_h are assigned to the first rank; all other subjects are pooled into the second rank. In particular, only high-ability subjects can reach the first rank; all low-ability subjects are ranked second (i.e., last), regardless of their effort. Under the *M ranking system*, given by the middle threshold alone, all subjects with high outcomes (H_h or H_l) are assigned to the first rank. Subjects of both ability types with low outcomes (L_h or L_l) are ranked second (last).

Under the *TM ranking system*, given by the top and middle thresholds, subjects achieving outcome H_h are ranked first, and those achieving H_l are ranked second; subjects achieving L_h or L_l are ranked last. Thus, both high- and low-ability subjects can improve their rank through effort, but only the former can be ranked first. Under the *MB ranking system*, given by the middle and bottom thresholds, all subjects with outcome H_h or H_l are ranked first; thus, both high- and low-ability subjects can reach the first rank. Subjects with outcome L_h are ranked second, and those with outcome L_l are ranked last.

Finally, under the *TMB ranking system*, which contains the top, middle, and bottom thresholds, each outcome

Table 1. (Color online) Ranking Systems Used in the Experiment

Ranking system	Description
T (top threshold)  	Subjects achieving H_h are ranked first. Those achieving H_l, L_h, or L_l are ranked second.
M (middle threshold)  	Subjects achieving H_h or H_l are ranked first. Those achieving L_h or L_l are ranked second.
TM (top and middle thresholds)  Rank 3 Rank 2 Rank 1 L_l L_h H_l H_h	Subjects achieving H_h are ranked first, those achieving H_l are ranked second, and those achieving L_h or L_l are ranked third.
MB (middle and bottom thresholds)  Rank 3 Rank 2 Rank 1 L_l L_h H_l H_h	Subjects achieving H_h or H_l are ranked first, those achieving L_h are ranked second, and those achieving L_l are ranked third.
TMB (top, middle, and bottom thresholds)  Rank 4 Rank 3 Rank 2 Rank 1 L_l L_h H_l H_h	Subjects achieving H_h are ranked first, those achieving H_l are ranked second, those achieving L_h are ranked third, and those achieving L_l are ranked fourth.

has its own rank. Subjects with H_h are ranked first, those with H_l second, those with L_h third, and those with L_l last. This ranking system is the most granular one possible in our setting; it provides full information about outcomes.[6]

3.4. Sample and Protocol

A total of 112 physicians and 240 medical students participated in our experiment. We applied the Main experimental treatment to all 112 physicians and 128 of the medical students and the Control treatment to the remaining 112 medical students.

The experiments were conducted in meeting and seminar rooms at hospitals and at the University of Cologne Faculty of Medicine. Tables equipped with laptops were arranged so that groups of four subjects could sit together. When entering the room, each subject drew a number indicating which laptop they would use. Subjects performed all tasks in the experiment anonymously at their computers. Only after finishing the first part did they receive instructions for the rest of the experiment. (See Section B.2 of the Online Appendix for the complete instructions.)

We used a random-choice payment technique. Each subject's payment was determined by single decisions drawn at random (i.e., one draw for each subject) from the first or third part of the experiment. After the experiment, we elicited subjects' altruism with an incentivized standard dictator game (Forsythe et al. 1994). The experimental sessions concluded with a questionnaire on the subjects' demographics. Each session lasted for about 45 minutes. The average payoff was about 17 euro for medical students and 54 euro for physicians. A total of about 9,854 euro was transferred to the Christoffel Blindenmission.

Table 2 summarizes the characteristics of our sample of subjects. Whereas our experimental design focuses mainly on within-subject comparisons, we observe that medical-student subjects are balanced across the Main and Control treatments.

3.5. Behavioral Predictions

We now translate Hypotheses 1 and 2 into predictions of behavior in the experiment. We first consider the effort choices of high- and low-ability subjects separately, then aggregate them.

For high-ability subjects, the top and middle thresholds *motivate*, whereas the bottom threshold *demotivates* (Hypothesis 1). When the top and bottom thresholds are both added to a ranking system, effort increases—that is, the motivating effect of the top threshold for a high-ability subject outweighs the demotivating effect of the bottom one (Hypothesis 2). Therefore, denoting the effort chosen by a high-ability subject i under a given ranking system by $e_{i,h}(\cdot)$, we expect $e_{i,h}(T) \leq e_{i,h}(TM)$ and $e_{i,h}(MB) \leq e_{i,h}(M) \leq e_{i,h}(TMB) \leq e_{i,h}(TM)$.

Table 2. Sample Characteristics

Subject pool Experimental treatment	All subjects ($N = 352$)	Physicians MAIN ($N = 112$)	Medical students MAIN ($N = 128$)	Medical students CONTROL ($N = 112$)
Age (in years)	27.19 (8.48)	35.96 (9.11)	22.91 (3.51)	23.13 (3.59)
Female	0.67 (0.46)	0.65 (0.48)	0.70 (0.45)	0.65 (0.48)
Clinical experience (in years)	–	10.10 (8.47)	–	–
Study term	–	–	4.69 (3.03)	4.76 (3.29)
Test score	4.31 (1.58)	3.87 (1.41)	4.70 (1.50)	4.32 (1.71)
Altruism	2.38 (1.49)	2.78 (1.49)	2.20 (1.42)	2.19 (1.51)

Notes. This table shows means, with standard deviations in parentheses. The test score is the number of correct answers on the nine-question ability assessment administered in the second stage of the experiment. The level of altruism is elicited in a simple dictator game, played after the main experiment, in which subjects allocate 4 ECU between themselves and a charity.

Similarly, for the low-ability subjects, the middle and bottom thresholds *motivate*, whereas the top threshold *demotivates* (Hypothesis 1). When both the top and bottom thresholds are added to a ranking system, effort increases (Hypothesis 2). Thus, defining $e_{i,l}(\cdot)$ analogously to $e_{i,h}(\cdot)$, we expect $e_{i,l}(T) \leq e_{i,l}(TM) \leq e_{i,l}(M) \leq e_{i,l}(TMB) \leq e_{i,l}(MB)$.

Furthermore, if we add the top (bottom) threshold to a ranking system, the motivating effect for high-ability (low-ability) subjects exceeds the demotivating effect for low-ability (high-ability) subjects (Hypothesis 2). Because our experiment has equal numbers of low- and high-ability subjects, we expect the average efforts $\overline{e}(\cdot)$ (the average of both ability types) to satisfy the following: $\overline{e}(M) \leq \overline{e}(TM) \leq \overline{e}(TMB)$ and $\overline{e}(M) \leq \overline{e}(MB) \leq \overline{e}(TMB)$.

4. Results

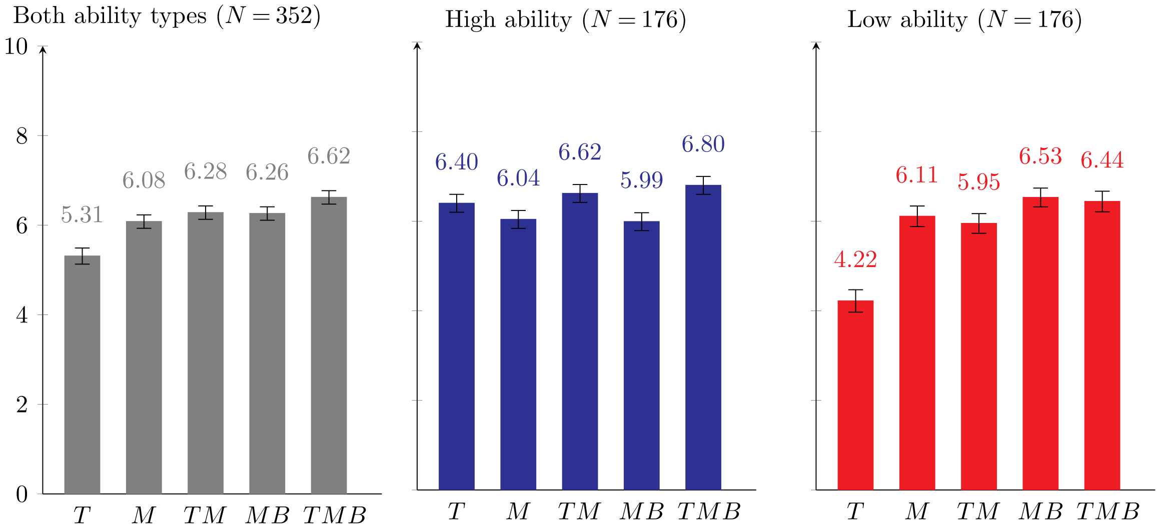
In this section, we first test our behavioral predictions using nonparametric statistics and then use parametric regressions to test the robustness of our main results (Section 4.1). We also compare the results under each ranking system to the non-ranking baseline (Section 4.2).

4.1. Comparison of Ranking Systems

Figure 2 summarizes the effort choices made in the experiment. The left panel shows the average effort across both ability types; the middle and right panels show the average efforts of the high- and low-ability subjects, respectively. We see that effort choices vary substantially, depending on the ranking system in use; however, the direction and intensity of this variation depend on the ability type.

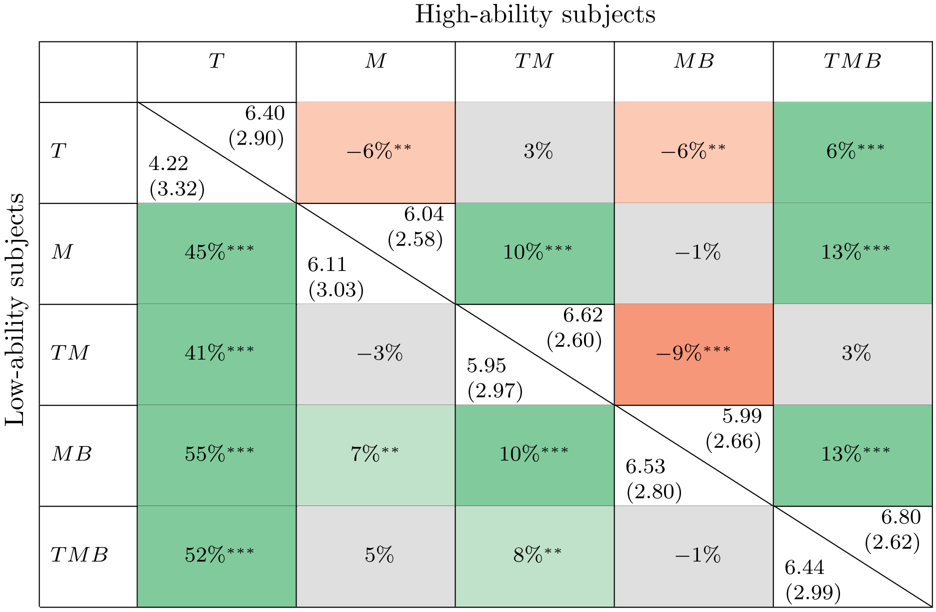
Figure 3 presents pairwise comparisons of the efforts in the five ranking systems, broken down by ability type; values for high-ability (low-ability) subjects are shown above (below) the diagonal. The white cells along the diagonal show the mean effort and standard deviation (in parentheses) for each ranking system. The value in each above-diagonal cell indicates the percentage change (in effort by high-ability subjects) if the

Figure 2. (Color online) Effort Under Each Ranking System



Note. This figure shows the average effort (with standard error bars) for the full population of subjects, the low-ability subjects, and the high-ability subjects under each ranking system.

Figure 3. (Color online) Pairwise Comparisons of Effort Choices by Ability Type



High-ability subjects

Low-ability subjects

	T	M	TM	MB	TMB
T	6.40 (2.90) 4.22 (3.32)	−6%**	3%	−6%**	6%***
M	45%***	6.04 (2.58) 6.11 (3.03)	10%***	−1%	13%***
TM	41%***	−3%	6.62 (2.60) 5.95 (2.97)	−9%***	3%
MB	55%***	7%**	10%***	5.99 (2.66) 6.53 (2.80)	13%***
TMB	52%***	5%	8%**	−1%	6.80 (2.62) 6.44 (2.99)

Notes. This table shows the relative difference between subjects' mean effort choices for each pair of ranking systems. Values for high-ability (low-ability) subjects are shown above (below) the diagonal. The value in each above-diagonal cell indicates the percentage change in effort if the ranking system given by the row of the cell (denoted by R_{row}) is replaced by the one given by the column (denoted by R_{col}); that is, it equals $\bar{e}_h(R_{\text{col}})/\bar{e}_h(R_{\text{row}}) - 1$. The value in each below-diagonal cell indicates the percentage change if R_{col} is replaced by R_{row}, that is, $\bar{e}_l(R_{\text{row}})/\bar{e}_l(R_{\text{col}}) - 1$. The cells on the diagonal report mean effort choices and standard deviations for high-ability and low-ability subjects. The p values are as follows: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$ (based on Holm-corrected Wilcoxon signed-rank tests for paired samples).

ranking system given by the row of the cell is replaced by the one given by the column. For instance, the −6% in the second cell of the first row means that the effort under M (second column) is 6% less than the effort under T (first row). The below-diagonal cells should be interpreted in the opposite way. For instance, the 45% in the first cell of the second row means that the effort under M (second row) is 45% greater than the effort under T (first column).

For high-ability subjects, the average efforts under the five ranking systems (which we denote by $\bar{e}_h(\cdot)$) can be ordered as follows: $\bar{e}_h(MB) < \bar{e}_h(M) < \bar{e}_h(T) < \bar{e}_h(TM) < \bar{e}_h(TMB)$. In particular, as hypothesized, the ranking systems that include the top threshold (T, TM, TMB) lead to significantly higher effort than those that do not (M, MB). On the other hand, adding the bottom threshold does not significantly affect effort. Indeed, effort is lower in MB than in M, whereas effort is larger in TMB and TM. Lastly, adding the middle threshold (going from T to TM) only insignificantly increases effort. In summary, the top threshold strongly motivates high-ability subjects (which one might interpret as first-place loving), whereas the middle threshold tends to only increase effort. Adding the bottom threshold implies insignificant mixed effects. The TMB ranking system induces the highest effort.

For low-ability subjects, the average efforts $\bar{e}_l(\cdot)$ can be ordered as follows: $\bar{e}_l(T) < \bar{e}_l(TM) < \bar{e}_l(M) < \bar{e}_l(TMB) < \bar{e}_l(MB)$. The T ranking system leads to significantly lower effort than the other four systems, all of which include at least one threshold achievable by low-ability subjects. (Specifically, M, TM, TB, and TMB all induce between 41% and 55% greater effort than T.) Adding the top threshold to an existing ranking system (going from M to TM or MB to TMB) causes an insignificant decrease in effort. Adding the bottom threshold, on the other hand (going from M to MB or TM to TMB), increases effort by between 7% and 8% (which one might interpret as last-place aversion). Finally, adding the middle threshold (going from T to TM) drastically increases effort. In summary, adding achievable thresholds always induces a significant increase in effort, whereas there is only suggestive evidence that adding the unachievable top threshold decreases effort. The ranking system consisting of only the single unachievable threshold (T) induces by far the least effort.

These results are mostly in line with our predictions and support the hypotheses in Section 2.2. We summarize as follows.

Result 1 (Ranking System Design and Ability). *A subject's effort level depends on the ranking system design and on the subject's ability type.*

a. ***Effort is increasing in the number of achievable thresholds in the ranking system.*** *The effort of high-ability subjects is highest under the ranking systems that include the two thresholds that they can reach, namely, TM* *and TMB. The effort of low-ability subjects is highest under the ranking systems that include the two thresholds that they can reach, namely, MB and TMB.*

b. ***The presence of a threshold that a subject cannot surpass tends to decrease that subject's effort.*** *Low-ability subjects exert the least effort under the ranking system T. Adding achievable thresholds reduces the negative impact of the unachievable top threshold.*

c. ***The presence of a threshold that a subject is guaranteed to surpass does not significantly affect that subject's effort.*** *High-ability individuals exhibit an insignificant decrease in effort under MB relative to M and an insignificant increase in effort under TMB relative to TM.*

We now examine aggregate effort without differentiating between ability types. Figure 4 presents pairwise comparisons of the average effort in the five ranking systems across the full sample (both ability types). We see that effort is 18% higher under *TM* than under *T*; this highlights the strongly motivating effect of adding a threshold (the middle threshold) that both ability types can reach. In addition, effort is 3% higher under both *TM* and *MB* than under *M*, a weakly significant increase; this suggests that adding the top (bottom) threshold motivates high-ability (low-ability) subjects more than it demotivates low-ability (high-ability) subjects, as proposed in Hypothesis 2. It is not clear whether this effect depends on the ability type, because the effort levels under *TM* and *MB* are equal.

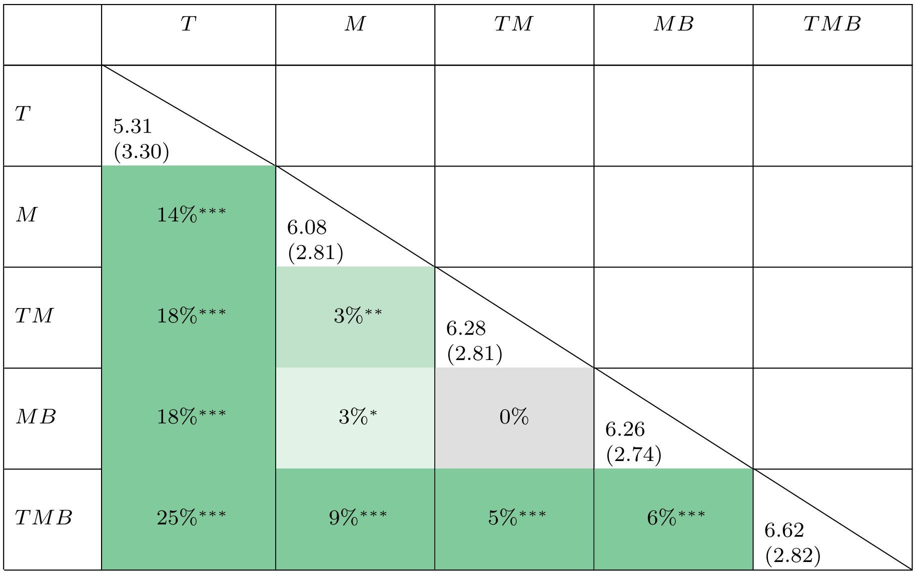
Next we observe that effort is 9% higher under *TMB* than under *M*. This supports the prediction that when we add both the top and bottom thresholds, the motivating effects of each new threshold for one ability type outweigh the potentially demotivating effects for the other type. (As we saw in Figure 3, the effort under *TMB* exceeds the effort under *M* by 13% for high-ability individuals and by 5% (although this value is insignificant) for low-ability individuals.) Also, the *TMB* ranking system induces greater effort, by a significant margin, than any other ranking system. These findings are in line with our predictions, which said that *TMB* not only should yield the highest aggregate effort but also should be at least the second-best ranking system for both ability types.

In contrast, the *T* ranking system induces significantly less effort—between 14% and 25% less—than any other ranking. In light of the disaggregated results (Figure 3), we infer that this result is driven by the strongly demotivating effect of the top threshold for low-ability subjects. These observations can be summarized as follows.

Result 2 (Salience of Extreme Thresholds). *Adding the middle threshold, which all subjects can reach, to a ranking system increases aggregate effort. Adding either the top or the bottom threshold also increases effort. The TMB ranking system induces more effort, and the T ranking system induces less effort, than any other ranking system.*

4.1.1. Robustness. We now check the robustness of the results described above. To analyze potential behavioral differences between physicians and medical students, as well as different health outcomes (from the Control treatment), while accounting for the influence of individual characteristics, we use the following

Figure 4. (Color online) Pairwise Comparisons of Effort Choices, Aggregated Across Ability Types



	T	M	TM	MB	TMB
T	5.31 (3.30)				
M	$14\%^{***}$	6.08 (2.81)			
TM	$18\%^{***}$	$3\%^{**}$	6.28 (2.81)		
MB	$18\%^{***}$	$3\%^{*}$	0%	6.26 (2.74)	
TMB	$25\%^{***}$	$9\%^{***}$	$5\%^{***}$	$6\%^{***}$	6.62 (2.82)

Notes. This figure shows the relative difference between subjects' mean effort choices for each pair of ranking systems, aggregated across both ability types. Values are calculated as in the below-diagonal cells of Figure 3. The cells on the diagonal report mean effort choices and standard deviations for all subjects. The *p* values are as follows: **p* < 0.1, ***p* < 0.05, ****p* < 0.01 (based on Holm-corrected Wilcoxon signed-rank tests for paired samples).

specification:

$$e_i = \beta_0 + \beta_\gamma \mathbf{R}_\gamma + \beta_1 \text{ALTRUISM}_i + \beta_2 \text{FEMALE}_i + \varepsilon_i,$$

where e_i is subject i's effort choice, β_0 is the intercept, $\mathbf{R}_\gamma$ is a column vector of dummies for the ranking systems, and β_γ is a row vector of the corresponding coefficients, with $\gamma \in \{M, TM, MB, TMB\}$ (here, the T ranking system serves as the reference category). The term ALTRUISM_i reflects subject i's altruistic concerns (as measured by the incentivized dictator game after the end of the main experiment), FEMALE_i is a dummy for subject i's gender, and ε_i is the error term.

The estimation results and Wald tests, as shown in Table 3, are consistent with the results of our nonparametric analysis. The results for Models (1)−(3) confirm that, in aggregate, the T ranking system always leads to significantly lower effort than any other system. In each specification, the two-threshold ranking systems always lead to higher effort than any single-threshold system, and *TMB* always leads to the highest effort. Because *TMB* always leads to higher effort than M (and the difference between the two sides is significant), these findings are in line with Result 2.

The estimation results and Wald test results for Models (4)−(6) confirm that adding the top threshold always significantly increases effort for high-ability subjects. Furthermore, adding the bottom threshold never significantly lowers effort. Models (7)−(9) confirm that low-ability subjects always exert significantly lower effort under T than under any other ranking system. For all specifications, we find that $\bar{e}_l(T) < \bar{e}_l(TM), \bar{e}_l(M) < \bar{e}_l(TMB) < \bar{e}_l(MB)$, which is in line with Result 1. All estimation results are robust to the inclusion of controls for gender and altruism.

4.2. Comparison of Effort Choices with Ranking to the Non-Ranking Baseline

We now analyze how effort choices in the presence of ranking compare with the non-ranking baseline. Table 4 provides descriptive statistics. In aggregate, the mean effort in the non-ranking baseline is $\bar{e}(\text{Base}) = 6.28$. Low-ability subjects expend significantly higher effort ($\bar{e}_l(\text{Base}) = 6.62$) than high-ability subjects ($\bar{e}_h(\text{Base}) = 5.94$) (Mann–Whitney U test, $p = 0.007$).[7]

Table 4 shows that introducing the *TMB* ranking system significantly increases effort (for both ability types

Table 3. Effects of Ranking System Designs on Effort, Relative to T

	Both ability types			High ability			Low ability		
Subject pool	Phys.	Stud.	Stud.	Phys.	Stud.	Stud.	Phys.	Stud.	Stud.
Exp. treatment	MAIN	MAIN	CONTROL	MAIN	MAIN	CONTROL	MAIN	MAIN	CONTROL
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
M	0.411*	0.945***	0.920***	−0.054	−0.562**	−0.429	0.875**	2.453***	2.268***
	(0.243)	(0.262)	(0.287)	(0.283)	(0.237)	(0.278)	(0.392)	(0.386)	(0.438)
TM	0.554***	1.016***	1.348***	0.071	−0.000	0.625***	1.036***	2.031***	2.071***
	(0.197)	(0.207)	(0.250)	(0.196)	(0.146)	(0.192)	(0.333)	(0.347)	(0.447)
MB	0.688***	1.000***	1.170***	0.071	−0.906***	−0.304	1.304***	2.906***	2.643***
	(0.233)	(0.303)	(0.305)	(0.266)	(0.295)	(0.291)	(0.371)	(0.410)	(0.462)
TMB	0.830***	1.469***	1.607***	0.536**	0.094	0.625**	1.125***	2.844***	2.589***
	(0.216)	(0.247)	(0.316)	(0.226)	(0.160)	(0.265)	(0.368)	(0.402)	(0.548)
Altruism	0.541***	0.406***	0.350**	0.546**	0.196	−0.143	0.614**	0.567**	0.932***
	(0.180)	(0.153)	(0.172)	(0.239)	(0.186)	(0.245)	(0.266)	(0.224)	(0.199)
Female	1.158**	0.206	−0.381	1.552**	−0.393	−0.130	0.760	0.738	−1.134*
	(0.514)	(0.433)	(0.560)	(0.695)	(0.501)	(0.728)	(0.763)	(0.613)	(0.677)
Constant	3.562***	4.380***	4.169***	4.203***	7.006***	5.627***	2.707***	1.928***	2.753***
	(0.715)	(0.580)	(0.503)	(0.991)	(0.602)	(0.652)	(0.994)	(0.742)	(0.631)
Differences between coefficients Wald tests of the following hypotheses H_0									
$M = TM$	−0.143	−0.070	−0.429**	−0.125	−0.562**	−1.054***	−0.161	0.422*	0.196
$M = MB$	−0.277*	−0.055	−0.250*	−0.125	0.344*	−0.125	−0.429	−0.453*	−0.375*
$M = TMB$	−0.420**	−0.523***	−0.688***	−0.589**	−0.656***	−1.054***	−0.250	−0.391	−0.321
$TM = MB$	−0.134	0.016	0.179	0.000	0.906***	0.929***	−0.268	−0.875***	−0.571**
$TM = TMB$	−0.277*	−0.453***	−0.259	−0.464**	−0.094	0.000	−0.089	−0.813***	−0.518
$MB = TMB$	−0.143	−0.469***	−0.437***	−0.464**	−1.000***	−0.929***	0.179	0.063	0.054
Observed decisions	560	640	560	280	320	280	280	320	280
Subjects	112	128	112	56	64	56	56	64	56
R^2	0.126	0.076	0.064	0.155	0.047	0.033	0.140	0.217	0.263

Notes. This table shows estimation results from ordinary least squares regressions with robust standard errors clustered at the individual subject level. The reference category is the T ranking system. "Female" is a gender dummy that equals 1 for female subjects and 0 for male subjects. Altruism was measured as described in Table 2. For a Tobit specification, see Table C.5 in Section C.3 of the Online Appendix. The p values are as follows: $^{*}p < 0.1$; $^{**}p < 0.05$; and $^{***}p < 0.01$.

Table 4. Effort Choices Under Rank Feedback vs. the Nonranking Baseline

	Both ability types		High ability		Low ability	
Ranking system	Mean (SD)	%-Diff to *Base*	Mean (SD)	%-Diff to *Base*	Mean (SD)	%-Diff to *Base*
Base line	6.28 (2.79)		5.94 (2.68)		6.62 (2.87)	
T	5.31 (3.30)	−15.43***	6.40 (2.90)	7.75***	4.22 (3.32)	−36.22***
M	6.08 (2.81)	−3.21	6.04 (2.58)	1.72	6.11 (3.03)	−7.64*
TM	6.28 (2.81)	0.09	6.62 (2.60)	11.48***	5.95 (2.97)	−10.13***
MB	6.26 (2.74)	−0.23	5.99 (2.66)	0.96	6.53 (2.80)	−1.29
TMB	6.62 (2.82)	5.43***	6.80 (2.62)	14.55***	6.44 (2.99)	−2.75

Notes. This table shows descriptive statistics for the subjects' baseline effort choices (in the absence of ranking) and their effort choices under each of the five ranking systems. In the baseline task, when subjects' ability types had not yet been determined, each subject made two choices, one as if they had low ability and one as if they had high ability. Their choices in the presence of ranking are compared with the baseline choice corresponding to their actual type. The changes in effort relative to the baseline are given in percentages. The p values are as follows: *$p < 0.1$; ***$p < 0.01$ (based on Holm-corrected Wilcoxon signed-rank tests for paired samples).

in aggregate) relative to the baseline, by about 5%. By contrast, introducing the *T* ranking system decreases effort by about 16% compared with the baseline. The use of the other ranking systems has no significant effect on aggregate effort. This is in line with the previous literature, which has shown that rank feedback can either increase or decrease effort.

Considering individual abilities provides a more nuanced view. On the one hand, high-ability subjects are never demotivated by any form of ranking. Particularly, they expend more effort under any ranking system than in the non-ranking baseline. Moreover, for all ranking systems that include the top threshold (i.e., *T*, *TM*, and *TMB*), which high-ability subjects can reach but low-ability subjects cannot, the increase in effort is significant and lies between 8% and 15%.

On the other hand, low-ability subjects are never motivated by any form of ranking. They expend less effort under any ranking system than in the non-ranking baseline. For the *T*, *M*, and *TM* ranking systems (the ones that contain only one achievable threshold, or none), the drop in effort relative to the baseline is significant; it is largest under *T*, about 36%.

5. Implications and Discussion

Our lab-in-the-field experiment sheds light on how the design of a ranking system, in conjunction with an individual physician's level of ability, affects effort provision in healthcare. In aggregate (for a team containing both low- and high-ability physicians), the largest performance improvements in response to rank feedback occur under a ranking system with multiple thresholds, spanning the entire range of possible outcomes. A threshold near the top of the range is necessary to motivate high-ability physicians to improve their effort. But lower thresholds are also needed to motivate low-ability physicians who have no chance of reaching the top rank. Intuitively, by providing low-ability physicians with attainable ranks to strive for, a clinical leader can offset the potential demotivating effects of unattainably high ranks.

For a given clinical team, the appropriate level for the topmost threshold will depend on the mix of abilities within the team and the goals of the clinical leader. The topmost threshold should be attainable for a significant portion of the team yet high enough to make the top rank fairly exclusive. (In particular, if a team has very few high-ability members, or if the clinical leader is focused on motivating low-ability individuals, then the topmost threshold should not be set too high.) As mentioned in the introduction, our findings may provide guidance on the design of feedback mechanisms related to a wide range of high-volume healthcare activities that admit performance measures at the individual-physician level. However, our results need to be interpreted in light of our specific experimental design and its limitations, which we discuss below.

5.1. Features of the Experimental Design

One might argue that the stated-effort method used in our experiment may not adequately capture the field setting and the psychological forces involved in exerting actual effort (Charness et al. 2018, p. 74); perhaps it would have been more appropriate to use a framed field experiment that included real-effort tasks resembling actual clinical work (see, e.g., Eilermann et al. 2019, Kim et al. 2020). There are good reasons, however, to prefer the stated-effort approach in our context. First, it removes any uncertainty regarding an individual's cost to exert a certain level of effort, which varies for real-effort tasks because of factors such as the individual's level of knowledge (see, e.g., Müller and Schotter 2010). Second, to address our hypotheses, it is important to distinguish between ability and effort. We assess subjects' abilities using a real-effort task; although the nature of ability in a clinical setting is admittedly much more complex, the task requires skills such that our ability assignments are not random. Third, our stated-effort tasks are less time-consuming than real-effort tasks, allowing us to include multiple tasks comparing a comprehensive range of ranking systems.

Furthermore, although our experimental setting is rather stylized, we have confirmed through interviews with clinical leaders ($N = 7$) in the areas of gastroenterology, orthopedic surgery, and pediatrics that our stated-effort task accurately captures the main incentives a physician faces when rendering health services in a clinical setting. All of the clinical leaders identified real-world activities from their respective clinical areas to which our rank-feedback approach could apply (e.g., lumbar punctures and appendectomies (in pediatrics), appendicitis detection (in pediatric radiology), and intravenous access placement (in emergency care)). Also, on a questionnaire, more than 80% of the physicians participating in the experiment indicated a clinical task that would resemble the stylized decision problem they had just faced. In all of the example activities listed here, the distinction between ability and effort is practically relevant because a higher level of ability (attained through experience or education) can help physicians achieve better outcomes while expending the same effort.[8] In the short run, physicians can change their effort levels but not their levels of ability, so relative performance feedback primarily affects effort provision.

The fact that we privately informed subjects of their ability types may have affected their subsequent effort choices. Murthy and Schafer (2011), for instance, showed that framed feedback can affect agents' allocation decisions. However, this step was unavoidable in our experiments because subjects needed to know their types in order know their achievable outcomes in the decision task, rather than their beliefs about their ability levels. We believe that type disclosure did not have a heterogeneous effect on subjects' decisions across various ranking systems because the mode of disclosure was the same for all subjects and all effort choices were made after type disclosure.

Finally, we considered rank feedback in peer groups of four subjects. The small group size enabled nonanonymity while still allowing for two ability types, each assigned to two subjects. Admittedly, real-world clinical teams typically are comprised of more than four physicians. However, for relative performance feedback whose impact is due solely to social comparison, evidence on the relevance of the group size is scarce. Some tournament studies include varied group sizes, but these studies have not found clear evidence of whether increasing the group size increases or decreases effort (Dechenaux et al. 2015).

5.2. Generalizability

Our results should extend to much broader applications than we could test within the confines of our experiment. Our targeted experimental design with physicians as subjects is meant to make our findings more directly applicable to the relevant population (Gneezy and Imas 2017, p. 440). However, in any setting where individuals care about status, as discussed in Section 2 and the model of status concern in the Online Appendix A, the essential effects of the choice of ranking system should be the same regardless of parameters such as the nature of the task, the exact outcome distribution, or the group size. What may change is the *magnitude* of these effects (particularly the effect of feedback relative to the baseline without rank feedback), because setting-specific parameters will determine the overall importance of individuals' status. Our robustness checks support this expectation; we find that the order of the ranking systems (in terms of their effects on effort, as described in Results 1 and 2) is robust to changes in the subject pool, the health outcome distribution, and other covariates (Table 3).

5.3. Replicability

Apart from the limitations mentioned above, one might argue that our main results may be difficult to replicate and that the relatively large number of compared ranking systems may have affected the behavioral results. However, our pilot experiment, which had a smaller subject pool ($N = 116$) and included seven ranking system designs (see Section 3.3), yielded patterns similar to those in Results 1 and 2. In the pilot, for the sake of feasibility, we divided the subject pool between two treatments, one covering three ranking systems, and the other four. Within each treatment, for subjects with high and low ability, we observed the same order of mean effort levels as in Result 1. In line with Result 2, *TMB* was the most attractive ranking system. For more on the pilot experiment, see Section C.1 of the Online Appendix.

6. Concluding Remarks

Taken together, our results provide clinical leaders with valuable insights into the design of performance-feedback mechanisms that could directly affect the delivery of care. The potential impact of these insights could be similar to adjustments of operational processes in work design (see, e.g., Song et al. 2015, Tucker 2016, Ibanez et al. 2018, Berry Jaeker and Tucker 2020). We anticipate, for example, that clinical leaders could apply our findings in providing structured relative performance feedback during individual performance review meetings with their clinical teams. It is important to note, however, that although rankings can be a useful tool, they may be most effective when integrated into a broader culture of continuous improvement and professional development. The emphasis should be on supporting clinicians in their efforts to provide high-quality care, rather than solely relying on competitive measures.

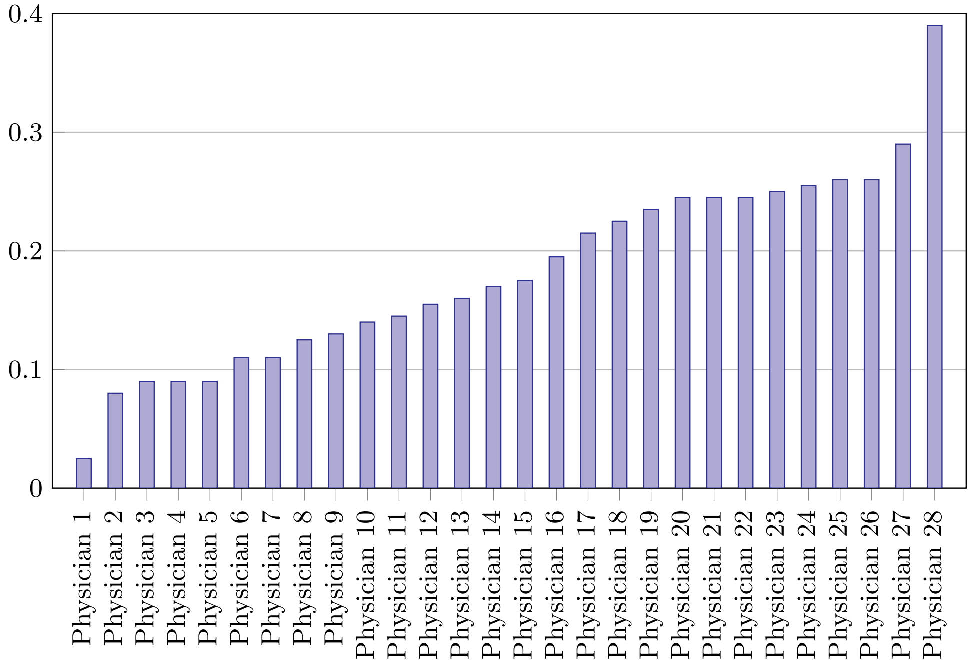
Our results also draw attention to the important challenge of how to design (non)monetary incentives that account for physician characteristics so as to improve the quality of care. An appealing feature of our experimental design is that it can readily be adapted to the study further factors affecting physician effort. We have thus introduced a valuable and easily scalable experimental paradigm for studying other factors that may play a role in the effectiveness of relative performance feedback among peers in a clinical setting.

Acknowledgments

The authors are grateful for numerous valuable comments and suggestions from Stefan Scholtes, the department editor, an anonymous associate editor, and four anonymous reviewers. The authors also gratefully acknowledge comments from Aurélien Baillon, Gary Bolton, Alex Chan, Yan Chen, Ben Greiner, Izabela Jelovac, Mathias Kifmann, Johanna Kokot, Ludwig Kuntz, Amol Navathe, Axel Ockenfels, Carol Propper, Lise Rochaix, Rupert Sausgruber, Matthias Sutter, and Achim Wambach as well as by participants of seminars and conferences at the BEH workshop Innsbruck, dggö Hamburg, dggö Augsburg, emlyon Business School, Gfew Erfurt, Imperial College Business School, Karlsruhe Institute of Technology, MPI Bonn, University of Cologne, University of Hamburg, VfS Berlin, WU Vienna, and ZEW Mannheim. The authors also gratefully acknowledge support in the recruiting of subjects, facilitating the experiments, and providing clinical insights to map our experimental design to Dr. med. Christoph Baltin, Professor Dr. med. Jörg Dötsch, Dr. med. Frank Eifinger, Professor Dr. med. Peer Eysel, Professor Dr. med. Tobias Goeser, Professor Dr. med. Christoph Härtel, Dr. Yassin Karary, Dr. med. Martin Maus, Professor Dr. med. Jessica Leers, and Professor Dr. med. Raymond Voltz. Emma Brauckhoff, Clara Lena Barrocou, Timon Brinker, Anthony Ge, Jonathan Goedeke, Rieke Heinrichs, Marina Kancheva, Ben Mrakovcic, Nils Ridder, Juliane Stracke, Till Stange, and Simon Weidtmann provided excellent research assistance.

Appendix. Illustrative Example

Figure A.1. (Color online) Adenoma Detection Rates from a Major UK Clinic – Excluding Bowel Cancer Screening Program (BCSP), Names Anonymized and Sorted from Lowest to Highest



Endnotes

[1] The adenoma detection rate is the proportion of screening colonoscopies performed by a physician that detect one or more adenomas (Corley et al. 2014). Figure A.1 in the appendix gives an example of the distribution of adenoma detection rates across a clinical team of 28 physicians in a major U.K. clinic.

[2] Examples of high-volume activities in other clinical areas are lumbar punctures and appendectomies in pediatrics, the detection of appendicitis in pediatric radiology, and intravenous access placement in emergency care. Further examples of relative performance feedback can be found in Song et al. (2018), which considers the length of stay of discharged high-acuity patients in a hospital emergency department, and Navathe et al. (2020), which focuses on primary-care services such as advance care planning, obesity control, cervical-cancer screening, childhood immunization, flu vaccination, and screening for clinical depression.

[3] With additional financial incentives, feedback often provides information to subjects about their chances of reaching the incentives. For an excellent review of the effects of relative performance information in contests, see Dechenaux et al. (2015). A noteworthy related stream of the literature considers a setting in which patients have access to public information on the performance of individual physicians, for instance, in cardiac surgery (see, e.g., Dranove et al. 2003, Cutler et al. 2004). This setting makes it possible to disentangle the effects of feedback from those of financial incentives as well as from demand-side effects (e.g., Kolstad 2013).

[4] Our choice of Cohen's *d* was based on the observed values from our pilot experiment; for more information, see Section C of the Online Appendix. Notice that in the preregistration document (AsPredicted No. 130723), we also proposed comparing effort choices in the presence of rank feedback to a non-ranking baseline (Section 4.2). The corresponding sample size thus accounts for even more pairwise comparisons.

[5] Analogous mechanisms to make concern for patient health salient have been used in other experiments on physician behavior (see, e.g., Hennig-Schmidt et al. 2011, Waibel and Wiesen 2021, Brosig-Koch et al. 2022). Similarly, to make individuals' stated choices salient, the framed field experiment of Chan (2023), which addressed patients' preferences in choosing healthcare providers, implemented a matching of hypothetical physician profiles to real physicians, who then actually rendered medical services.

[6] Two other potential ranking systems exist: one containing only the bottom threshold (*B*) and one with the top and bottom thresholds (*TB*). However, our pilot experiment, which included all seven possible systems, suggested that these two were less relevant in practice; see Section C of the Online Appendix. Furthermore, the *B* system, which singles out the worst performers, seems inappropriate for small clinical teams (whereas *T* and *M* may be reasonable). The *TMB*, *TM*, and *MB* systems all contain two achievable thresholds for at least one of the ability types; this is not true of *TB*. Because a smaller set of ranking systems would be easier for subjects to make sense of and compare during the experiment, we omitted *B* and *TB* from consideration.

[7] This difference is driven by within-subject differences rather than between-subject differences in effort choices. In the first stage of the experiment, before their types were tested, 46.6% of the subjects chose higher effort when supposing that they had low ability than when supposing that they had high ability, whereas 11.9% did the reverse. We further find that the subject's actual type (as realized in the second stage of the experiment) does not affect their two baseline (non-ranking) effort choices (Wilcoxon signed-rank test, $p = 0.5989$ and $p = 0.9288$).

[8] For example, high-ability physicians may consistently achieve successful intravenous access placements with minimal effort and minimal discomfort to the patient thanks to, for example, their experience, fine motor skills, or anatomical understanding. A physician with less ability may need to invest more effort into each placement (e.g., by preparing more extensively or being more diligent during the insertion) to achieve similar levels of success.

References

Ashraf N, Bandiera O, Lee SS (2014) Awards unbundled: Evidence from a natural field experiment. *J. Econom. Behav. Organ.* 100(C): 44–63.

Azmat G, Iriberri N (2010) The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *J. Public Econom.* 94(7–8):435–452.

Bandiera O, Barankay I, Rasul I (2013) Team incentives: Evidence from a firm level experiment. *J. Eur. Econom. Assoc.* 11(5):1079–1114.

Berry Jaeker JA, Tucker AL (2020) The value of process friction: The role of justification in reducing medical costs. *J. Ops. Management* 66(1–2):12–34.

Besley T, Ghatak M (2008) Status incentives. *Amer. Econom. Rev.* 98(2):206–211.

Bradler C, Dur R, Neckermann S, Non A (2016) Employee recognition and performance: A field experiment. *Management Sci.* 62(11):3085–3099.

Brosig-Koch J, Griebenow M, Kifmann M, Then F (2022) Rewards for information provision in patient referrals: A theoretical model and an experimental test. *J. Health Econom.* 86:102677.

Brown DJ, Ferris DL, Heller D, Keeping LM (2007) Antecedents and consequences of the frequency of upward and downward social comparisons at work. *Organ. Behav. Hum. Decis. Proc.* 102(1):59–75.

Buell RW (2021) Last-place aversion in queues. *Management Sci.* 67(3):1430–1452.

Chan A (2023) Discrimination against doctors: A field experiment. Technical report, Harvard Business School.

Chan DC, Gentzkow M, Yu C (2022) Selection with variation in diagnostic skill: Evidence from radiologists. *Quart J. Econom.* 137(2):729–783.

Charness G, Gneezy U, Henderson A (2018) Experimental methods: Measuring effort in economics experiments. *J. Econom. Behav. Organ.* 149(C):74–87.

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

Chen DL, Schonger M, Wickens C (2016) oTree—An open-source platform for laboratory, online, and field experiments. *J. Behav. Exp. Finance* 9(C):88–97.

Coffman K, Klinowski D (2025) Gender and preferences for performance feedback. *Management Sci.* 71(4):3497–3516.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, Hillsdale, NJ).

Corley DA, Jensen CD, Marks AR, Zhao WK, Lee JK, Doubeni CA, Zauber AG, et al. (2014) Adenoma detection rate and risk of colorectal cancer and death. *New Engl. J. Med.* 370(14):1298–1306.

Cotofan M (2021) Learning from praise: Evidence from a field experiment with teachers. *J. Public Econom.* 204(C):104540.

Cutler DM, Huckman RS, Landrum MB (2004) The role of information in medical markets: An analysis of publicly reported outcomes in cardiac surgery. *Amer. Econom. Rev.* 94(2):342–346.

Dechenaux E, Kovenock D, Sheremeta RM (2015) A survey of experimental research on contests, all-pay auctions and tournaments. *Exp. Econom.* 18(4):609–669.

Dranove D, Kessler D, McClellan M, Satterthwaite M (2003) Is more information better? The effects of "report cards" on health care providers. *J. Political Econom.* 111(3):555–588.

Dubey P, Geanakoplos J (2010) Grading exams: 100,99,98,… or A,B,C? *Games Econom. Behav.* 69(1):72–94.

Edelman B, Larkin I (2015) Social comparisons and deception across workplace hierarchies: Field and experimental evidence. *Organ. Sci.* 26(1):78–98.

Eilermann K, Halstenberg K, Kuntz L, Martakis K, Roth B, Wiesen D (2019) The effect of expert feedback on antibiotic prescribing in pediatrics: Experimental evidence. *Med. Decision Making* 39(7):781–795.

Ericsson KA, Charness N (1994) Expert performance: Its structure and acquisition. *Amer. Psychologist* 49(8):725–747.

Forsythe R, Horowitz JL, Savin NE, Sefton M (1994) Fairness in simple bargaining experiments. *Games Econom. Behav.* 6(3):347–369.

Gerhards L, Siemer N (2016) The impact of public and private feedback on worker performance—Evidence from the lab. *Econom. Inquiry* 54(2):1188–1201.

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

Gneezy U, Imas A (2017) Chapter 10—Lab in the field: Measuring preferences in the wild. Banerjee AV, Duflo E, eds. *Handbook of Field Experiments*, Handbook of Economic Field Experiments, vol. 1 (North-Holland, Amsterdam), 439–464.

Gowrisankaran G, Joiner K, Léger PT (2023) Physician practice style and healthcare costs: Evidence from emergency departments. *Management Sci.* 69(6):3202–3219.

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

Hannan RL, McPhee GP, Newman AH, Tafkov ID (2013) The effect of relative performance information on performance and effort allocation in a multi-task environment. *Accounting Rev.* 88(2): 553–575.

Hennig-Schmidt H, Selten R, Wiesen D (2011) How payment systems affect physicians' provision behaviour—An experimental investigation. *J. Health Econom.* 30(4):637–646.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Sci.* 64(9):4389–4407.

Kim SH, Tong J, Peden C (2020) Admission control biases in hospital unit capacity management: How occupancy information hurdles and decision noise impact utilization. *Management Sci.* 66(11):5151–5170.

Kolstad JT (2013) Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *Amer. Econom. Rev.* 103(7):2875–2910.

Kuhnen CM, Tymula A (2012) Feedback, self-esteem, and performance in organizations. *Management Sci.* 58(1):94–113.

Kuziemko I, Buell RW, Reich T, Norton MI (2014) "Last-place aversion": Evidence and redistributive implications. *Quart. J. Econom.* 129(1):105–149.

Loch CH, Wu Y (2008) Social preferences and supply chain performance: An experimental study. *Management Sci.* 54(11):1835–1849.

Moldovanu B, Sela A, Shi X (2007) Contests for status. *J. Political Econom.* 115(2):338–363.

Müller W, Schotter A (2010) Workaholics and dropouts in organizations. *J. Eur. Econom. Assoc.* 8(4):717–743.

Murthy US, Schafer BA (2011) The effects of relative performance information and framed information systems feedback on performance in a production task. *J. Inform. Systems* 25(1): 159–184.

Navathe AS, Volpp KG, Bond AM, Linn KA, Caldarella KL, Troxel AB, Zhu J, et al. (2020) Assessing the effectiveness of peer comparisons as a way to improve the health care quality: Examining whether peer comparisons feedback provided to primary care providers may impact quality of care. *Health Affairs* 39(5):852–861.

Newman AH, Tafkov ID (2014) Relative performance information in tournaments with different prize structures. *Accounting Organ. Soc.* 39(5):348–361.

Niewoehner RJ, Staats BR (2022) Focusing provider attention: An empirical examination of incentives and feedback in flu vaccinations. *Management Sci.* 68(5):3680–3702.

Rege M, Telle K (2004) The impact of social approval and framing on cooperation in public good situations. *J. Public Econom.* 88(7):1625–1644.

Roels G, Su X (2014) Optimal design of social comparison effects: Setting reference groups and reference points. *Management Sci.* 60(3):606–627.

Schnieder C (2022) How relative performance information affects employee behavior: A systematic review of empirical research. *J. Accounting Literature* 44(1):72–107.

Selten R (1965) *Die Strategiemethode Zur Erforschung Des Eingeschränkt Rationalen Verhaltens im Rahmen Eines Oligopolexperimentes*, Beiträge zur experimentellen Wirtschaftsforschung (J.C.B. Mohr (Paul Siebeck), Tübingen), 136–168.

Siau K, Green JT, Hawkes ND, Broughton R, Feeney M, Dunckley P, Barton JR, Stebbing J, Thomas-Gibson S (2019) Impact of the Joint Advisory Group on Gastrointestinal Endoscopy (JAG) on endoscopy services in the UK and beyond. *Frontline Gastroenterol.* 10(2):93–106.

Singh M, Zureich J (2024) Do physicians improve more from positive or negative feedback? *Management Sci.* 71(5):4198–4222.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Sci.* 61(12):3032–3053.

Song H, Tucker AL, Murrell KL, Vinson DR (2018) Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Sci.* 64(6):2628–2649.

Suls J, Wheeler L (2000) *A Selective History of Classic and Neo-Social Comparison Theory* (Springer US, Boston), 3–19.

Tafkov ID (2013) Private and public relative performance information under different compensation contracts. *Accounting Rev.* 88(1):327–350.

Tucker AL (2016) The impact of workaround difficulty on frontline employees' response to operational failures: A laboratory experiment on medication administration. *Management Sci.* 62(4): 1124–1144.

Turkoglu A, Tucker A (2022) The demotivating effects of relative performance feedback on middle-ranked workers' performance. Research Paper, Boston University Questrom School of Business (4242303).

Valori R, Cortas G, De Lange T, Balfaqih OS, de Pater M, Eisendrath P, Falt P, et al. (2018) Performance measures for endoscopy services: A European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *Endoscopy* 50(12):1186–1204.

Waibel C, Wiesen D (2021) An experiment on referrals in health care. *Eur. Econom. Rev.* 131:103612.

Zizzo DJ (2002) Between utility and cognition: The neurobiology of relative position. *J. Econom. Behav. Organ.* 48(1):71–91.