

How the Design of Ranking Systems and Ability Affect Physician Effort*

Katharina Huesmann, Yero Samuel Ndiaye, Christian Waibel, Daniel Wiesen

December 2024

Abstract

While relative performance feedback in the form of rankings appears to be effective in improving healthcare outcomes, it may have either motivating or demotivating effects for individual physicians. Potential factors influencing such effects include a physician's level of 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—i.e., 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, while 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

*Katharina Huesmann: University of Münster; e-mail: Katharina.Huesmann@wiwi.uni-muenster.de. Yero Samuel Ndiaye: University of Cologne and Max Planck Institute; e-mail: ndiaye@wiso.uni-koeln.de. Christian Waibel: ETH Zurich; e-mail: cwaibel@ethz.ch. Daniel Wiesen: University of Cologne; e-mail: wiesen@wiso.uni-koeln.de. We are grateful for valuable comments and suggestions from Gary Bolton, Alex Chan, Yan Chen, Ben Greiner, 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, Gfew Erfurt, Imperial Business School, Karlsruhe Institute of Technology, MPI in Bonn, University of Cologne, University of Hamburg, WU Vienna, and ZEW Mannheim. We thank Emma Brauckhoff, Clara Barrocou, Timon Brinker, Anthony Ge, Jonathan Goedeke, Rieke Heinrichs, Marina Kancheva, Ben Mrakovcic, Nils Ridder, Julianne Stracke, Till Stange and Simon Weidtmann for their excellent research assistance. We also gratefully acknowledge the support in 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, Professor Dr. med. Jessica Leers, Professor Dr. med. Raymond Voltz. Financial support from the Center for Social and Economics Behavior (C-SEB) at the University of Cologne and from the University of Münster (grant for junior researchers) is gratefully acknowledged. Parts of this research was conducted while Daniel Wiesen was employed at the Institute of Health and Society at the Department of Health Management and Health Economics, University of Oslo supported by the Research Council of Norway (IRECOHEX, grant-no.: 231776). Yero Samuel Ndiaye gratefully acknowledges support by the German Science Foundation through Germany's Excellence Strategy (EXC 2126/1 390838866).

provide rank feedback using a system under which physicians of all ability types can improve their rank through increased effort.

Key words: Ranking design, peer feedback, lab-in-the-field experiment, status concerns, ability

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 (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 (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 (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 healthcare 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 non-trivial 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, while those who have no chance of reaching it may be demotivated by their failure to do so. Since physicians vary in ability, this creates a potential trade-off: 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 (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 UK’s Joint Advisory Group on Gastrointestinal Endoscopy,

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

²Examples of high-volume activities in other clinical areas are lumbar punctures and appendectomies in pedi-

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 motivation to the vast majority of physicians, since 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 (e.g., Schnieder 2022; for more details, see Section 2). We use a well-powered and pre-registered 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 lab paradigm but is conducted in a naturalistic setting (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, as 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 trade-offs 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

atrics, 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.

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 the subjects' choices.

In our stylized set-up, 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, while 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% to 25% higher than those under the other systems. The system consisting of only the top threshold yields the lowest effort levels, 13% to 20% less than the other ranking systems.

Compared to 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 to 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 to the baseline, while 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) show 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) report that relative performance feedback improves quality in primary-care organizations. Niewoehner and Staats (2022) find 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 (e.g., Hannan et al. 2013, Tafkov 2013, Gerhards and Siemer 2016), while private feedback may do so by stimulating people's self-image concerns (e.g., Tafkov 2013, Gill et al. 2019). Kuhnen and Tymula (2012) observe 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 (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 (2023) show 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) report that private feedback improves performance, but they find 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 Iribarri 2010, Kuziemko et al. 2014, Gill et al. 2019, Niewoehner and Staats 2022). Correspondingly, Turkoglu and Tucker (2022) find that receiving feedback causes the performance of middle-ranked individuals to suffer. On the other hand, Bradler et al. (2016) report 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 (Buell 2021, Müller and Schotter 2010, Cotofan 2021). These studies generally consider 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

³With 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 (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).

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 2024). In addition, since 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 her performance in a clinical activity—i.e., 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 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, while 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 (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 (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), since 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 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 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 threshold, effort decreases.*
- (c) *For individuals who cannot reach outcomes below the threshold, effort decreases.*

Since 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 trade-offs 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 (e.g., Müller and Schotter 2010, Azmat and Iribarri 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. We therefore expect that motivating effects of these thresholds are higher than 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, 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 Subsection 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 to 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 pre-registered 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 e-mails 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 e-mail 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 non-parametric 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 $e \in \{0, 1, 2, \dots, 10\}$, 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 $L_l < L_h < H_l < H_h$. (For an illustration, see Figure 1.) For either type, the probability of a high outcome (H_h or H_l) is $P_H(e)$ 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 $1 - P_H(e)$. 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 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 $H_h = 25$.

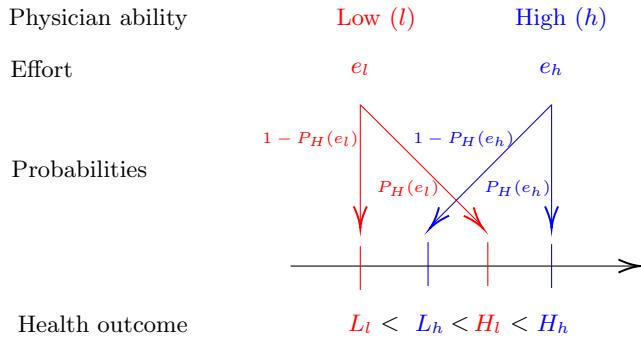
Although the tasks in the experiment deal with abstract patients, we incorporated the factor of a physician's concern for real patients by translating the health benefits from each subject's decision into monetary terms and transferring this amount to the Christoffel Blindenmission, a charitable organization,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

⁴Our choice of Cohen's d was based on the observed values from our pilot experiment; for more information, see Section C.1 of the online appendix. Notice that in the pre-registration 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.

⁵Analogous mechanisms to make concern for patient health salient have been used in other experiments on physician behavior (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.

Figure 1: 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 $P_H(e_l)$ and outcome L_l with probability $1 - P_H(e_l)$. If a high-ability subject chooses effort e_h , they achieve outcome H_h with probability $P_H(e_h)$ and outcome L_h with probability $1 - P_H(e_h)$.

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 their ability type was high and one assuming 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 their ability type and of the outcomes that they 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 Subsection 3.3). We used a one-shot decision set-up, 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 Subsection 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 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 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

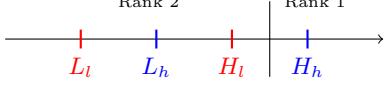
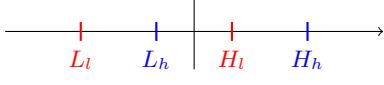
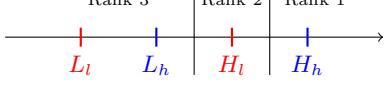
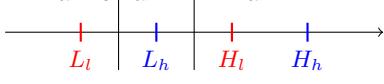
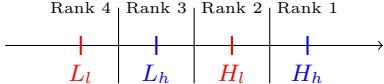
3.4 Sample and Protocol

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 indicat-

⁶Two 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.1 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 . Since 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.

Table 1: 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) 	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) 	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) 	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.

ing 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. While our experimental design focuses mainly on within-subject comparisons, we observe that medical-student subjects are balanced across the MAIN and CONTROL treatments.

Table 2: Sample characteristics

Subject pool: Experimental treatment:	All subjects (<i>N</i> = 352)	Physicians MAIN (<i>N</i> = 112)	Medical students MAIN (<i>N</i> = 128)	Medical students CONTROL (<i>N</i> = 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.

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*, while 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 total effort chosen by the high-ability subjects under a given ranking system by $e_{i,h}(\cdot)$, we expect $e_{i,h}(T) \leq e_{i,h}(TM) \leq e_{i,h}(MB) \leq e_{i,h}(M) \leq e_{i,h}(TMB) \leq e_{i,h}(TM)$.

Similarly, for the low-ability subjects, the middle and bottom thresholds *motivate*, while the top threshold *demotivates* (Hypothesis 1). When the top and bottom thresholds are both 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). Since our experiment has equal numbers of low- and high-ability subjects, we expect the aggregate efforts $e(\cdot)$ (the totals across both ability types) to satisfy the following: $e(M) \leq e(TM) \leq e(TMB) \leq e(MB) \leq e(TM)$.

4 Results

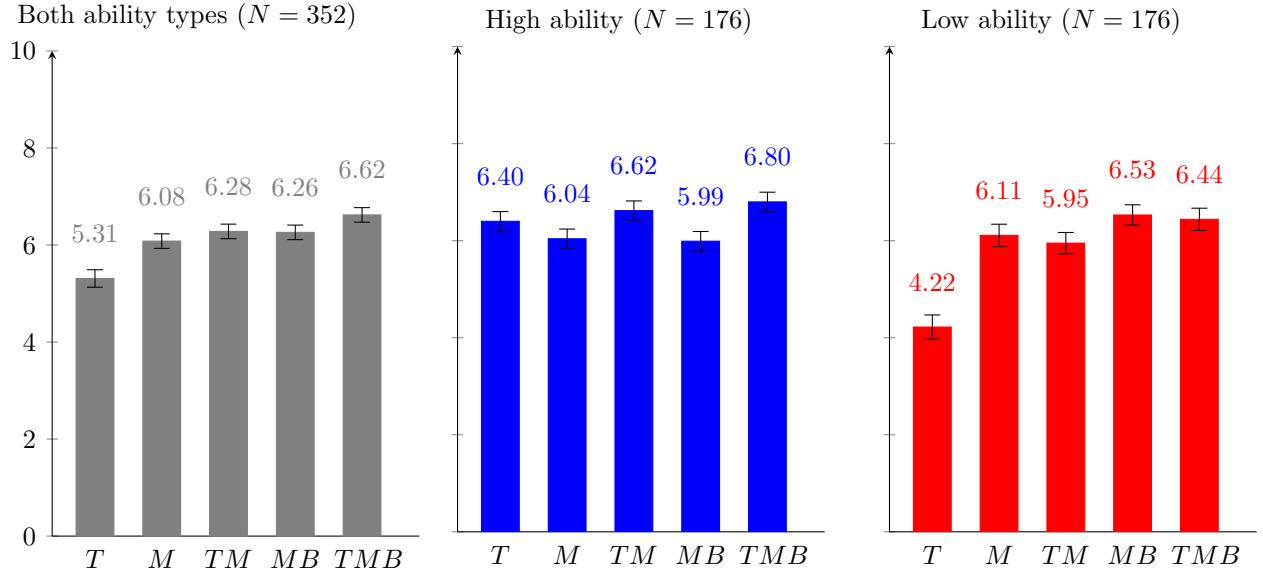
In this section, we first test our behavioral predictions using non-parametric statistics, then use parametric regressions to test the robustness of our main results (Subsection 4.1). We also compare the results under each ranking system to the non-ranking baseline (Subsection 4.2).

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.

Table 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 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).

Figure 2: Effort under each ranking system



Notes. 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.

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 while

effort is larger in TMB and TM . Lastly, adding the middle threshold (going from T to TM) only insignificantly increases effort. In sum, the top threshold strongly motivates high-ability subjects (which one might interpret as first-place loving), while the middle threshold only tends to 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 induces 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 sum, adding achievable thresholds always induces a significant increase in effort, while 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.

Table 3: Pairwise comparisons of effort choices, by ability type

High-ability subjects

	T	M	TM	MB	TMB
T	6.40 (2.90) 4.22 (3.32)	-6%**	3%	-6%**	6%***
M	45%*** 6.11 (3.03)	6.04 (2.58)	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} , i.e., $\bar{e}_l(R_{\text{row}})/\bar{e}_l(R_{\text{col}}) - 1$. 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).

These results are mostly in line with our predictions and support the hypotheses in Subsec-

tion 2.2. We summarize as follows:

Result 1 (Ranking system design and ability) *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. Table 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, since 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 Table 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 (Table 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*

Table 4: Pairwise comparisons of effort choices, aggregated across ability types

	<i>T</i>	<i>M</i>	<i>TM</i>	<i>MB</i>	<i>TMB</i>
<i>T</i>	5.31 (3.30)				
<i>M</i>	14%***	6.08 (2.81)			
<i>TM</i>	18%***	3%**	6.28 (2.81)		
<i>MB</i>	18%***	3%*	0%	6.26 (2.74)	
<i>TMB</i>	25%***	9%***	5%***	6%***	6.62 (2.82)

Notes. This table 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 Table 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).

threshold also increases effort. The TMB ranking system induces more effort, and the T ranking system induces less effort, than any other ranking system.

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 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}_γ 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 5, are consistent with the results of our non-parametric analysis. The results for Models (1) to (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. As 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.

Table 5: Effects of ranking system designs on effort, relative to T

Subject pool: Exp. treatment: Model:	Both ability types				High ability				Low ability			
	Phys. MAIN (1)	Stud. MAIN (2)	Stud. CONTROL (3)	Phys. MAIN (4)	Stud. MAIN (5)	Stud. CONTROL (6)	Phys. MAIN (7)	Stud. MAIN (8)	Stud. CONTROL (9)			
M	0.411* (0.243)	0.945*** (0.262)	0.920*** (0.287)	-0.054 (0.283)	-0.562** (0.237)	-0.429 (0.278)	0.875** (0.392)	2.453*** (0.386)	2.268*** (0.438)			
TM	0.554*** (0.197)	1.016*** (0.207)	1.348*** (0.250)	0.071 (0.196)	0.000 (0.146)	0.625*** (0.192)	1.036*** (0.333)	2.031*** (0.347)	2.071*** (0.447)			
MB	0.688*** (0.233)	1.000*** (0.303)	1.170*** (0.305)	0.071 (0.266)	-0.906*** (0.295)	-0.304 (0.291)	1.304*** (0.371)	2.906*** (0.410)	2.643*** (0.462)			
TMB	0.830*** (0.216)	1.469*** (0.247)	1.607*** (0.316)	0.536** (0.226)	0.094 (0.160)	0.625** (0.265)	1.125*** (0.368)	2.844*** (0.402)	2.589*** (0.548)			
Altruism	0.541*** (0.180)	0.406*** (0.153)	0.350*** (0.172)	0.546** (0.239)	0.196 (0.186)	-0.143 (0.245)	0.614** (0.266)	0.567** (0.224)	0.932*** (0.199)			
Female	1.158** (0.514)	0.206 (0.433)	-0.381 (0.560)	1.552** (0.695)	-0.393 (0.501)	-0.130 (0.728)	0.760 (0.763)	0.738 (0.613)	-1.134* (0.677)			
Constant	3.562*** (0.715)	4.380*** (0.580)	4.169*** (0.503)	4.203*** (0.991)	7.006*** (0.602)	5.627*** (0.652)	2.707*** (0.742)	1.928*** (0.742)	2.753*** (0.631)			
<i>Differences between coefficients</i>												
<i>Wald tests of the following hypotheses H_0</i>												
$M = TM$	-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***	-0.588***	-0.589**	-0.656***	-1.054***	-0.250	-0.391	-0.321			
$TM = MB$	-0.134	0.016	0.179	0.000	0.904***	0.929***	-0.268	-0.875***	-0.571*			
$TM = TMB$	-0.277*	-0.453***	-0.259*	-0.464**	-0.464**	0.000	-0.089	-0.813***	-0.518			
$MB = TMB$	-0.143	-0.469***	-0.437***	-0.437***	-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 which 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The estimation results and Wald test results for Models (4) to (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) to (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 to the non-ranking baseline. Table 6 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 6 shows that introducing the TMB ranking system significantly increases effort (for both ability types in aggregate) relative to the baseline, by about 5%. By contrast, introducing the T ranking system decreases effort by about 16% compared to 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.

Table 6: Effort choices under rank feedback versus the non-ranking baseline

Both ability types		High ability		Low ability	
Mean (s.d.)	%-Diff to Base	Mean (s.d.)	%-Diff to Base	Mean (s.d.)	%-Diff to Base
<i>Non-ranking baseline:</i>					
Base	6.28 (2.79)	5.94 (2.68)		6.62 (2.87)	
<i>Ranking systems:</i>					
T	5.31 (3.30)	-15.43***	6.40 (2.90)	7.75***	4.22 (3.32)
M	6.08 (2.81)	-3.21	6.04 (2.58)	1.72	6.11 (3.03)
TM	6.28 (2.81)	0.09	6.62 (2.60)	11.48***	5.95 (2.97)
MB	6.26 (2.74)	-0.23	5.99 (2.66)	0.96	6.53 (2.80)
TMB	6.62 (2.82)	5.43***	6.80 (2.62)	14.55***	6.44 (2.99)

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 to 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: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ (based on Holm-corrected Wilcoxon signed-rank tests for paired samples).

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

⁷This 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 they had low ability than when supposing 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 $p = 0.9288$).

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.

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 (e.g., Kim et al. 2020, Eilermann et al. 2019). 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 (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, while our experimental setting is rather stylized, we have confirmed through interviews with clinical leaders ($N = 7$) in the areas of gastroenterology, orthopaedic 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, since 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, show that framed feedback can affect agents' allocation decisions. However, this step was unavoidable in our experiments, as 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, as 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 non-anonymity 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).

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

⁸For example, high-ability physicians may consistently achieve successful intravenous access placements with minimal effort and minimal discomfort to the patient thanks to, e.g., 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.

(particularly the effect of feedback relative to the baseline without rank feedback), since 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 Result 2) is robust to changes in the subject pool, the health outcome distribution, and other covariates (Table 5).

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 Subsection 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 (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 while 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.

References

- Ashraf N, Bandiera O, Lee SS (2014) Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization* 100:44–63.
- Azmat G, Iribarri N (2010) The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics* 94(7-8):435–452.
- Bandiera O, Barankay I, Rasul I (2013) Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association* 11(5):1079–1114.
- Berry Jaeker JA, Tucker AL (2020) The value of process friction: The role of justification in reducing medical costs. *Journal of Operations Management* 66(1-2):12–34.
- Besley T, Ghatak M (2008) Status incentives. *American Economic Review* 98(2):206–211.
- Bradler C, Dur R, Neckermann S, Non A (2016) Employee recognition and performance: A field experiment. *Management Science* 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. *Journal of Health Economics* 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. *Organizational Behavior and Human Decision Processes* 102(1):59–75.
- Buell RW (2021) Last-place aversion in queues. *Management Science* 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. *Quarterly Journal of Economics* 137(2):729–783.
- Charness G, Gneezy U, Henderson A (2018) Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization* 149:74–87.
- Charness G, Masclet D, Villeval MC (2014) The dark side of competition for status. *Management Science* 60(1):38–55.
- Chen DL, Schonger M, Wickens C (2016) oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.
- Coffman K, Klinowski D (2024) Gender and preferences for performance feedback. *Management Science*.
- Cohen J (1988) *Statistical power analysis for the behavioral sciences* ((2nd ed.) Hillsdale, NJ: Erlbaum).
- Corley DA, Jensen CD, Marks AR, Zhao WK, Lee JK, Doubeni CA, Zauber AG, de Boer J, Fireman BH, Schottinger JE, Quinn VP, Ghai NR, Levin TR, Quesenberry CP (2014) Adenoma detection rate and risk of colorectal cancer and death. *New England Journal of Medicine* 370(14):1298–1306.
- Cotofan M (2021) Learning from praise: Evidence from a field experiment with teachers. *Journal of Public Economics* 204:104540.
- Cutler DM, Huckman RS, Landrum MB (2004) The role of information in medical markets: An analysis of publicly reported outcomes in cardiac surgery. *American Economic Review* 94(2):342–346.
- Dechenaux E, Kovenock D, Sheremeta RM (2015) A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics* 18:609–669.
- Dranove D, Kessler D, McClellan M, Satterthwaite M (2003) Is more information better? The effects of “report cards” on health care providers. *Journal of Political Economy* 111(3):555–588.

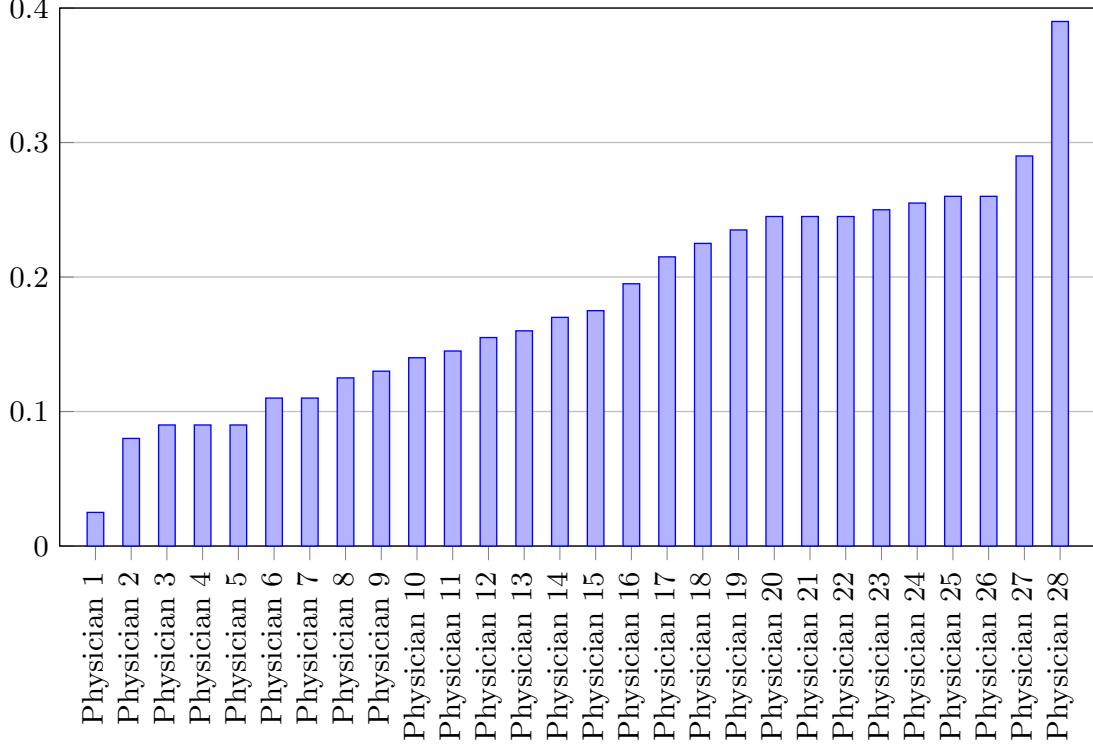
- Dubey P, Geanakoplos J (2010) Grading exams: 100,99,98,... or A,B,C? *Games and Economic Behavior* 69(1):72–94.
- Edelman B, Larkin I (2015) Social comparisons and deception across workplace hierarchies: Field and experimental evidence. *Organization Science* 26(1):78–98.
- Eilermann K, Halstenberg K, Kuntz L, Martakis K, Roth B, Wiesen D (2019) The effect of expert feedback on antibiotic prescribing in pediatrics: Experimental evidence. *Medical Decision Making* 39(7):781–795.
- Ericsson KA, Charness N (1994) Expert performance: Its structure and acquisition. *American Psychologist* 49(8):725.
- Forsythe R, Horowitz JL, Savin NE, Sefton M (1994) Fairness in simple bargaining experiments. *Games and Economic Behavior* 6(3):347–369.
- Gerhards L, Siemer N (2016) The impact of public and private feedback on worker performance – evidence from the lab. *Economic 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 Science* 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*, volume 1 of *Handbook of Economic Field Experiments*, 439–464 (North-Holland).
- Gowrisankaran G, Joiner K, Léger PT (2023) Physician practice style and healthcare costs: Evidence from emergency departments. *Management Science* 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. *The Accounting Review* 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. *The Accounting Review* 88(2):553–575.
- Hennig-Schmidt H, Selten R, Wiesen D (2011) How payment systems affect physicians' provision behaviour—an experimental investigation. *Journal of Health Economics* 30(4):637–646.
- Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 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 Science* 66(11):5151–5170.
- Kolstad JT (2013) Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review* 103(7):2875–2910.
- Kuhnen CM, Tymula A (2012) Feedback, self-esteem, and performance in organizations. *Management Science* 58(1):94–113.
- Kuziemko I, Buell RW, Reich T, Norton MI (2014) “last-place aversion”: Evidence and redistributive implications. *The Quarterly Journal of Economics* 129(1):105–149.
- Loch CH, Wu Y (2008) Social preferences and supply chain performance: An experimental study. *Management Science* 54(11):1835–1849.
- Moldovanu B, Sela A, Shi X (2007) Contests for status. *Journal of Political Economy* 115(2):338–363.
- Müller W, Schotter A (2010) Workaholics and dropouts in organizations. *Journal of the European Economic Association* 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. *Journal of Information Systems* 25(1):159–184.
- Navathe AS, Volpp KG, Bond AM, Linn KA, Caldarella KL, Troxel AB, Zhu J, Yang L, Matloubieh SE, Drye EE, 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, Organizations and Society* 39(5):348–361.
- Niewoehner RJ, Staats BR (2022) Focusing provider attention: An empirical examination of incentives and feedback in flu vaccinations. *Management Science* 68(5):3680–3702.
- Rege M, Telle K (2004) The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics* 88(7):1625–1644.
- Roels G, Su X (2014) Optimal design of social comparison effects: Setting reference groups and reference points. *Management Science* 60(3):606–627.
- Schnieder C (2022) How relative performance information affects employee behavior: A systematic review of empirical research. *Journal of Accounting Literature* 44(1):72–107.
- Selten R (1965) Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. 136–168, Beiträge zur experimentellen Wirtschaftsforschung, Tübingen: J.C.B. Mohr (Paul Siebeck).
- 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 Gastroenterology* 10(2):93–106.
- Singh M, Zureich J (2023) Do physicians improve more from positive or negative feedback? *Mimeo* .
- Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 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 Science* 64(6):2628–2649.
- Suls J, Wheeler L (2000) *A Selective History of Classic and Neo-Social Comparison Theory*, 3–19 (Boston, MA: Springer US).
- Tafkov ID (2013) Private and public relative performance information under different compensation contracts. *The Accounting Review* 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 Science* 62(4):1124–1144.
- Turkoglu A, Tucker A (2022) The demotivating effects of relative performance feedback on middle-ranked workers' performance. *Boston University Questrom School of Business Research Paper* (4242303).
- Valori R, Cortas G, De Lange T, Balfaqih OS, de Pater M, Eisendrath P, Falt P, Koruk I, Ono A, Rustemović N, 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. *European Economic Review* 131:103612.

Zizzo DJ (2002) Between utility and cognition: The neurobiology of relative position. *Journal of Economic Behavior & Organization* 48(1):71–91.

Appendix: Illustrative example

Figure Appendix.1: Adenoma detection rate - excluding Bowel Cancer Screening Program (BCSP), names anonymized and sorted from lowest to highest



A Model of status concerns

To study how the design of a ranking system affects effort choices, we build a simple model of status concerns. In our setting, each physician i on a team N of physicians chooses an effort level $e_i \in E \subset R_+$ at a cost of $c(e_i) \in R_+$, which is increasing in e_i . The effort e_i yields a stochastic patient benefit $x_i(e_i) \in X_i \subset R_+$. For simplicity, we assume that $X_i = \{L_i, H_i\}$ with $L_i < H_i$. The probability $P_H(e_i)$ of realizing the higher outcome(patient benefit) H_i is a strictly increasing function of effort e_i . The set X_i reflects the *ability* of physician i .

Based on the outcomes they realize, the physicians on the team are ranked according to some ranking scheme γ . For its formal definition we follow the definition of grading maps in Dubey and Geanakoplos (2010). Let \mathcal{R} denote the set of all possible orderings of N (with ties allowed). A *ranking system* is a map $\gamma : R_+^N \rightarrow \mathcal{R}$ which ranks physicians according to $\gamma(x)$ when $x = (x_i)_{i \in N}$ are the outcomes of the physicians; $\gamma_i(x) \in N$ is called the rank of physician i . All physicians with the same outcome are assigned to the same rank. We concentrate on absolute rankings, in which a physician's rank does not depend on other physicians' outcomes. (Dubey and Geanakoplos (2010) show why absolute rankings are superior to relative ones in a similar setting.) Therefore, we denote

$\gamma_i(x_i) = \gamma_i(x)$. Ranking schemes are monotone increasing, meaning that the higher a physician's outcome, the (weakly) higher her rank, with the convention that the first rank (i.e., $\gamma_i(x) = 1$) is the highest. More specifically, we consider rankings that are generated by a set of k thresholds $(t_j)_{j=1,\dots,k}$ with $t_1 > \dots > t_k$ that partition R_+ into $k + 1$ intervals which represent the outcomes that are bundled to a rank. More specifically, for $m = 1, \dots, k+1$ we have $\gamma_i(x_i) = m$ (i.e., physician i is assigned to rank m) if and only if $t_m \leq x_i < t_{m-1}$ with $t_{k+1} = 0$ and $t_0 = \infty$. As an example, all physicians with an outcome of at least t_1 are ranked first, all physicians with an outcome below t_k are ranked least on rank $k + 1$.

Given that physician i exerts effort e_i and the other physicians' outcomes are given by x_{-i} , we can define the following utility for physician i under a ranking scheme γ :

$$u_i^\gamma(e_i, x_{-i}) = \underbrace{\alpha_i(x_i(e_i))}_{\text{altruistic utility}} - \underbrace{c(e_i)}_{\text{effort cost}} + \underbrace{\lambda_i S^\gamma(x_i(e_i), x_{-i})}_{\text{status utility}}.$$

Here $\alpha_i(x_i(e_i))$ is the altruistic utility that physician i derives from generating the outcome $x_i(e_i)$, and $\lambda_i S^\gamma(x_i(e_i), x_{-i}) \in R$ is the utility that physician i derives from status under ranking scheme γ if the vector of all physicians' outcomes is $x = (x_i)_{i \in N}$. The coefficient $\lambda_i \in R_+$ captures the importance of status to physician i . Note that the outcome a physician achieves does not yield any monetary payoff; it is utility-relevant only through altruism and status concerns. Our key assumption on a physician's status utility is that it is increasing in the set of physicians ranked below her and decreasing in the set of physicians ranked above her. More formally, for a ranking scheme γ and outcomes $x = (x_i)_{i \in N}$, the status utility of physician i can be written as

$$S^\gamma(x_i, x_{-i}) = S(U^\gamma(x_i, x_{-i}), O^\gamma(x_i, x_{-i})),$$

where $U^\gamma(x_i, x_{-i}) = \{j | \gamma_j(x) > \gamma_i(x)\}$ denotes the set of physicians ranked below physician i , and $O^\gamma(x_i, x_{-i}) = \{j | \gamma_j(x) < \gamma_i(x)\}$ denotes the set of physicians ranked above physician i . The utility S^γ is assumed to be increasing in U^γ and decreasing in O^γ : For $U' \subset U$, we have $S(U', O) < S(U, O)$, and for $O' \subset O$, we have $S(U, O') > S(U, O)$. Given x_{-i} , the optimal effort e_i^γ for physician i is given by

$$e_i^\gamma = \arg \max E u_i^\gamma(e_i, x_{-i}).$$

To ensure that there is always a unique optimal effort, we assume $E u_i^\gamma(e_i, x_{-i})$ is concave in e_i .

Our definition of status utility is similar to the definitions used by Dubey and Geanakoplos (2010) and Moldovanu et al. (2007), although they assume that $S(U^\gamma, O^\gamma) = |U^\gamma| - |O^\gamma|$. By contrast, we make no specific assumptions on S^γ beyond monotonicity. Furthermore, unlike in Dubey and Geanakoplos (2010) and Moldovanu et al. (2007), utility is not derived purely from status; the physicians are altruistic as well.

Our aim is to study the impact on optimal effort if a ranking δ is made more granular through the addition of a threshold $t^* \in R_+$. The new ranking γ is then such that $\gamma_i(x_i) = \delta_i(x_i) + 1$ for all $x_i < t^*$ and $\gamma_i(x_i) = \delta_i(x_i)$ for all $x_i \geq t^*$. For an illustration, see Figure Appendix.2.

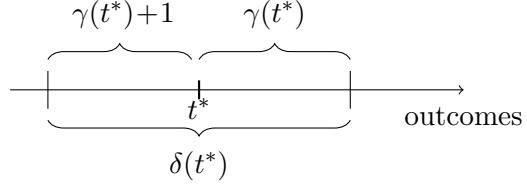


Figure Appendix.2: Adding a threshold t^*

The effect of the threshold t^* on a physician's status utility depends on whether the physician's outcome is above or below t^* . The addition of the threshold increases the status utility of outcomes above t^* , since it further separates such outcomes from those ranked lower. On the other hand, it decreases the status utility of outcomes below t^* , since it prevents the pooling of such outcomes with outcomes above t^* . Thus, whether the addition of t^* increases or decreases a physician's effort depends on whether the physician is able to reach outcomes above t^* . The proposition

below formalizes this and provides the theoretical foundation for Hypothesis 1 in Section 2.

Proposition 1 *Consider any physician i with binary stochastic outcomes $L_i < H_i$. Fix the outcomes x_{-i} of the other physicians. Given a ranking scheme δ , consider the ranking scheme γ formed by adding a new threshold $t^* \in R_+$ to δ . Then the optimal effort choices e_i^δ and e_i^γ under δ and γ satisfy the following:*

- If $L_i \leq t^* \leq H_i$, then $e_i^\gamma \geq e_i^\delta$.
- If $H_i < t^*$, then $e_i^\gamma \leq e_i^\delta$.
- If $L_i > t^*$, then $e_i^\gamma \leq e_i^\delta$.

We prove this proposition at the end of the section. According to Proposition 1, the effect on a physician of adding a threshold is closely linked to the physician's ability. A physician is most effectively motivated by thresholds that fall within the range of outcomes she can achieve; outside of that range, the ranking should optimally be coarse. Our interpretation of ability as the set of achievable outcomes is similar to that of Dubey and Geanakoplos (2010), whereas in Moldovanu et al. (2007), ability is linked to effort costs. Both works, however, emphasize the importance of ability to the design of a ranking scheme. While Dubey and Geanakoplos (2010) mainly assume complete information about ability distributions, Moldovanu et al. (2007) concentrate on incomplete information.

If physicians within a team differ in ability, thresholds near the ends of the range of outcomes achievable by the team may be motivating for some team members and demotivating for others. In Dubey and Geanakoplos (2010) and Moldovanu et al. (2007), the assumption $S(U^\gamma, O^\gamma) = |U^\gamma| - |O^\gamma|$ implies that the increase in status utility caused by the presence of a lower-ranked physician is equal to the decrease in status utility caused by the presence of a higher-ranked physician; that is, motivating effects do only depend on the absolute number of individuals

ranked above or below. In contrast, in our setting, trade-offs can be heterogeneous; for instance, the motivation to reach the first rank can be higher than to escape the last rank because being ranked higher than high performers might be more attractive compared to being ranked higher than low performers. In the experiment, we test for such salience effects by considering the addition of extremely low or high thresholds (see Hypothesis 2).

Proof of Proposition 1. Physician i and the outcome vector x_{-i} are fixed throughout the proof, so we omit the index i and the term x_{-i} in the notation. Let e^γ be physician i 's optimal effort under the ranking scheme γ ; it maximizes

$$Eu^\gamma(e) = E(\alpha(x(e))) - c(e) + \lambda ES^\gamma(x(e)). \quad (1)$$

e^δ is defined analogously. Now, first assume that $L \leq t^* \leq H$ and consider $e < e^\delta$. We show that

$$Eu^\gamma(x(e^\delta)) - Eu^\gamma(x(e)) \geq Eu^\delta(x(e^\delta)) - Eu^\delta(x(e)). \quad (2)$$

By assumption, $u^\gamma(x(e))$ is increasing in x . The right hand side is weakly greater than zero (by definition of e^δ). The inequality (2) then implies that the left hand side is also weakly greater than zero which implies $e^\gamma \geq e^\delta$. Therefore, part (i) of the proposition holds. To prove (2), we first substitute (1) into (2):

$$\begin{aligned} & Eu^\gamma(x(e^\delta)) - Eu^\gamma(x(e)) \geq Eu^\delta(x(e^\delta)) - Eu^\delta(x(e)) \\ \Leftrightarrow & ES^\gamma(x(e^\delta)) - ES^\gamma(x(e)) \geq ES^\delta(x(e^\delta)) - ES^\delta(x(e)) \\ \Leftrightarrow & P_H(e^\delta)S^\gamma(H) + (1 - P_H(e^\delta))S^\gamma(L) - P_H(e)S^\gamma(H) - (1 - P_H(e))S^\gamma(L) \\ & \geq P_H(e^\delta)S^\delta(H) + (1 - P_H(e^\delta))S^\delta(L) - P_H(e)S^\delta(H) - (1 - P_H(e))S^\delta(L) \\ \Leftrightarrow & [P_H(e^\delta) - P_H(e)][S^\gamma(H) - S^\delta(H)] \geq [P_H(e^\delta) - P_H(e)][S^\gamma(L) - S^\delta(L)] \\ \Leftrightarrow & S^\gamma(H) - S^\delta(H) \geq S^\gamma(L) - S^\delta(L). \end{aligned}$$

The assumption $L \leq t^* \leq H$ implies that $S^\gamma(H) - S^\delta(H) \geq 0$ and $S^\gamma(L) - S^\delta(L) \leq 0$. This is because, if H is realized, then the addition of the threshold t^* increases the status utility, since (weakly) more physicians are now ranked lower than without this threshold, while the set of outcomes ranked higher is unchanged. If L is realized, then the status utility decreases, since more outcomes are now ranked higher while the set of outcomes ranked lower is unchanged. Therefore (2) holds.

Next, assume that $t^* < L$ or $t^* > H$ and consider $e > e^\delta$. If (2) holds for these parameters, then $e^\gamma \leq e^\delta$, which yields parts (ii) and (iii) of the proposition. By transformations analogous to

those used above, $e > e^\delta$ implies that

$$\begin{aligned} Eu^\gamma(x(e^\delta)) - Eu^\gamma(x(e)) &\geq Eu^\delta(x(e^\delta)) - Eu^\delta(x(e)) \\ \Leftrightarrow S^\gamma(H) - S^\delta(H) &\leq S^\gamma(L) - S^\delta(L). \end{aligned}$$

If, under δ , the outcomes H and L are ranked the same, then $S^\gamma(H) = S^\gamma(L)$ and $S^\delta(H) = S^\delta(L)$ hold. This implies $S^\gamma(H) - S^\delta(H) = S^\gamma(L) - S^\delta(L)$, and therefore (2) holds. If, under δ , the outcomes H and L are ranked differently, then we look at $t^* > H$ and $t^* < L$ separately. For $t^* > H$, $S^\gamma(L) = S^\delta(L)$ (the sets of outcomes ranked higher and lower are unchanged) and $S^\gamma(H) - S^\delta(H) \leq 0$ (adding t^* increases only the set of outcomes ranked higher). For $t^* < L$, $S^\gamma(H) = S^\delta(H)$ and $S^\gamma(L) - S^\delta(L) \geq 0$. Therefore, $S^\gamma(H) - S^\delta(H) \leq S^\gamma(L) - S^\delta(L)$, which implies (2).

Note that we kept the outcomes x_{-i} of the other physicians fixed. Therefore, the optimal effort choices are not necessarily equilibrium choices. However, given potential cognitive restrictions around internalizing other physicians' effort changes, we consider this a reasonable approximation. Furthermore, if we alternatively define status utility based on the set of possible outcomes below or above one's individual rank (instead of the set of physicians), then physician i 's optimal effort choice is independent of the other physicians' effort choices and is therefore an equilibrium choice.

B Experiment materials

B.1 Experiment parameters

Table B.1: Effort levels, effort costs, and stochastic health outcomes in the experiment

Effort, e	Effort costs, $c(e)$	Low ability (l)		High ability (h)	
		Probability of patient benefit $L_l = 0$	Probability of patient benefit $H_l = 20$	Probability of patient benefit being $L_h = 5$	Probability of patient benefit $H_h = 25$
0	0	100%	0%	100%	0%
1	2	90%	10%	90%	10%
2	4	80%	20%	80%	20%
3	6	70%	30%	70%	30%
4	8	60%	40%	60%	40%
5	10	50%	50%	50%	50%
6	12	40%	60%	40%	60%
7	14	30%	70%	30%	70%
8	16	20%	80%	20%	80%
9	18	10%	90%	10%	90%
10	20	0%	100%	0%	100%

B.2 Instructions

Instructions for Part I of the Experiment

Thank you very much for participating in our decision-making experiment! Please read this experiment description carefully.

General Information

If you have any questions, please raise your hand. We will then come to you and answer your questions. Please do not talk to other participants until the end of the experiment.

For showing up on time, you will receive 20 euros.[5 euros for medical students.] During the experiment, you can earn more money. The amount of your earnings depends on your decisions. All decisions are made anonymously.

The payoff at the end of the experiment is anonymous, meaning no other participant knows about your payoff. The earnings during the experiment are displayed in Experimental Currency Units (ECU).

The experiment consists of three parts. **One decision** from the first or the third part will be **randomly** selected at the end of the experiment. The selected decision then determines your payoff. The amount from this randomly selected decision will be paid out to you at the following exchange rate:

$$1 \text{ ECU} = 3 \text{ EUR} [0.80 \text{ EUR for medical students}]$$

The experiment lasts about 45 minutes. [60 minutes for medical students.] At the end of the experiment, we ask you to answer some questions.

You make decisions in the experiment in the **role of a physician**.

Decision Situation

Each physician decides how much effort she wants to put in on a scale from 0 to 10 in treating the patient. Depending on your effort and the abilities allocated to you, the patient benefits differently from the treatment. The determination of abilities will be explained in more detail in the second part of the experiment.

Physicians with low abilities can achieve a benefit of 0 ECU [10 ECU in the CONTROL treatment] or 20 ECU for their patients. Physicians with high abilities can achieve a benefit of 5 ECU [15 ECU in the CONTROL treatment] or 25 ECU. The more a physician with low abilities makes an effort, the more likely it is that patients will have the benefit of 20 ECU from the treatment. The more a physician with high abilities makes an effort, the more likely it is that patients will have the benefit of 25 ECU from the treatment. The table below shows the exact relationship between the physician's abilities, the choice of effort level, and the patient benefit.

The effort in treating patients causes costs for the physician. These costs are independent of the physician's abilities. The more effort a physician makes, the higher the costs are. The costs for the physician for the different levels of effort are shown in the following table.⁹

Level of effort	Cost of effort (in ECU)	Physicians with low ability		Physicians with high ability	
		Probability of patient benefit = 0	Probability of patient benefit = 20	Probability of patient benefit = 5	Probability of patient benefit = 25
0	0	100%	0%	100%	0%
1	2	90%	10%	90%	10%
2	4	80%	20%	80%	20%
3	6	70%	30%	70%	30%
4	8	60%	40%	60%	40%
5	10	50%	50%	50%	50%
6	12	40%	60%	40%	60%
7	14	30%	70%	30%	70%
8	16	20%	80%	20%	80%
9	18	10%	90%	10%	90%
10	20	0%	100%	0%	100%

First Part of the Experiment

In the first part of the experiment, **you make two decisions**. You choose your level of effort for the treatment of a patient in the case that you have low abilities and in the case that you have high abilities. The other physicians do not receive any information about your decisions or the realized benefits of the patients.

⁹In the CONTROL treatment: $L_l=10$, $H_l=20$, $L_h=15$, and $H_h=25$.

Patients' Benefit

In the experiment, no participants in the role of patients are present in the lab. The amount that a patient receives from your treatment in the randomly selected decision will benefit a real patient. The amount will be donated to the Christoffel Blindenmission Deutschland e.V., 64625 Bensheim, which will use it to enable the treatment of patients with cataract. The transfer of the amount to the Christoffel Blindenmission Deutschland e.V. will be confirmed by a donation receipt. After the experiment, you can request a copy of the donation receipt via the e-mail address ndiaye@wiso.uni-koeln.de.

Physician's Profit

The physician's profit per decision is as follows:

$$\text{Revenue (in ECU)} = 20 - \text{Effort cost}$$

Payoff for physicians from the First Part of the Experiment

The payoff for physicians consists of the 20 euros [5 euros for medical students] for showing-up. This is in addition to the revenue (see previous section) from a randomly selected decision from this or the third part of the experiment (from a total of 7 decisions).

Payoff for Patients from the First Part of the Experiment

If a decision from the first part of the experiment is randomly chosen for payoff, the payoff for the patient corresponds to the benefit from the randomly selected decision.

Instructions for Part II and III of the Experiment

Second part of the experiment

In the second part, you answer nine questions as a physician. The number of questions you answer correctly determines whether you will play the role of a low-ability or high-ability physician in the third part of the experiment. The higher the number of questions you answer correctly is, the greater is the probability that you will be assigned the role of a high-ability physician in the third part of the experiment (more on this below). Other participants never see how many correct answers you have given.

Third part of the experiment

At the beginning of the third part of the experiment, please enter your first name. You will then be assigned to a group with three other physicians. You will remain assigned to the group at your table for the third part of the experiment. After all participants have entered their first names, they will be asked to stand up in groups and say their first names out loud in the order shown on the screen. To ensure that you know which participant is meant, even if they have the same first name, you will also be identified by your number within the group. This means that you are displayed as *participant_group_number*. The first names are only there so that people in the group can identify each other. Your first name will be overwritten with a random number at the end of the experiment, and the data will be analyzed anonymously.

The two participants in your group who answer the most questions correctly in the second part of the experiment will be assigned the role of a high-ability physician. The two participants who answered the fewest questions correctly in the second part will be assigned the role of a low-ability physician. If two participants have answered the same number of questions correctly, the order is randomized. At the beginning of the third part, each participant is shown his or her own ability (but not that of the others).

Decision-making situation and ranking of physicians within a group

In the third part of the experiment, **you make five decisions: You choose your level of effort for the treatment of each of the five patients**. After you have made your decisions, one of these decisions is selected at random.

For the selected decision, you will be ranked by name within your group based on the benefit achieved by the patient. The ranking is visible to all four physicians within your group, so that everyone in the group knows which group member has achieved which rank. The form of the ranking (i.e., which achieved benefits are assigned to which rank) depends on which of the five decisions is chosen at random. In contrast to your achieved rank, the other participants do not receive any information about your chosen effort level. Only you personally will receive an overview of your effort level and the patient's benefit at the end of the experiment.¹⁰

¹⁰The values mentioned correspond to the MAIN treatment and not the CONTROL treatment. In the instructions for the CONTROL treatment, they were adjusted accordingly.

When the first patient is treated, the **Ranking according to Type A** takes place if this decision is selected at random.

Ranking of type A:

Ranking A	
Rank 1	25
Rank 2	20
	5
	0

There are two ranks: Physicians whose patients have derived the benefit of 25 ECU from the treatment are rank 1; physicians whose patients have derived the benefit of 20, 5, or 0 ECU from the treatment are rank 2.

When treating the second patient, the **ranking according to type B** takes place if this decision is selected at random.

Ranking of type B:

Ranking B	
Rank 1	25
Rank 2	20
Rank 3	5
	0

There are three ranks: Physicians whose patients have derived the benefit of 25 ECU from the treatment are rank 1. Physicians whose patients have derived the benefit of 20 ECU from the treatment are rank 2. Physicians whose patients have derived the benefit of 5 or 0 ECU from the treatment are rank 3.

For the treatment of the third patient, the **ranking according to type C** takes place if this decision is selected at random.

Ranking of type C

Ranking C	
Rank 1	25
	20
Rank 2	5
	0

There are three ranks: Physicians whose patients have derived the benefit of 25 or 20 ECU from the treatment are rank 1. Physicians whose patients have derived the benefit of 5 ECU from the treatment are rank 2. Physicians whose patients have derived the benefit of 0 ECU from the treatment are rank 3.

For the treatment of the fourth patient, the **Ranking according to type D** takes place if this decision is selected at random.

Ranking of type D:

Ranking D	
Rank 1	25 20
<hr/>	
Rank 2	5 0

There are two ranks: Physicians whose patients have received the benefit of 25 or 20 ECU from the treatment are rank 1, and physicians whose patients have received the benefit of 5 or 0 ECU from the treatment are rank 2.

For the treatment of the fifth patient, the **ranking according to type E** takes place if this decision is selected at random. **Ranking of type E:**

Ranking E	
Rank 1	25
<hr/>	
Rank 2	20
<hr/>	
Rank 3	5
<hr/>	
Rank 4	0

There are four ranks: Physicians whose patients have benefited from 25 ECU are rank 1. Physicians whose patients have benefited from 20 ECU are rank 2. physicians whose patients have received the benefit of 5 ECU from the treatment are rank 3. physicians whose patients have received the benefit of 0 ECU from the treatment are rank 4.

[Each ranking was on a separate page in the original instructions.]

Payoff for physicians from the first or third part of the experiment

You will receive 20 euros [5 euros for medical students] for showing up. From the first and third part of the experiment, **a decision is randomly selected at the end of the experiment, and this decision determines your payoff.**

Payoff for patients from the first or third part of the experiment

The randomly selected decision from the first or third part of the experiment also determines the patients' payoff. This corresponds to the benefit realised by the patients in this decision.

B.3 Comprehension questions

B.3.1 Part I

1. Please assume you are a *high*-ability physician and choose an effort level of 0 when treating a patient.
 - a. What benefit (in ECU) does the patient gain from your treatment? *Correct answer: 5.*
 - b. What profit (in ECU) do you receive from choosing an effort level of 0? *Correct answer: 20.*
2. Please assume you are a *low*-ability physician and choose an effort level of 7 when treating a patient.
 - a. What is the probability (in %) that the patient benefit is 20 ECU? *Correct answer: 70.*
 - b. What is the probability (in %) that the patient benefit is 0 ECU? *Correct answer: 30.*
 - c. What profit do you make from the treatment? *Correct answer: 6.*
3. Assume you are a *high*-ability physician and choose an effort level of 6 when treating a patient.
 - a. What is the probability (in %) that the patient benefit is 25 ECU? *Correct answer: 60.*
 - b. What is the probability (in %) that the patient benefit is 5 ECU? *Correct answer: 40.*
4. In the first part of the experiment, do other physicians see what benefit the patient receives from your treatment? *Correct answer: No.*

B.3.2 Part III

5. Please assume you are a *low*-ability physician and your patient received a benefit of 20 ECU from your treatment. What is your rank in the respective ranking systems?
 - a. Ranking A: *Correct answer: 2*
 - b. Ranking B: *Correct answer: 2*

- c. Ranking C: *Correct answer: 1*
 - d. Ranking D: *Correct answer: 1*
 - e. Ranking D: *Correct answer: 2*
6. Please assume you are a *high*-ability physician and your patient received a benefit of 5 ECU from your treatment. What is your rank in the respective ranking systems?
- a. Ranking A: *Correct answer: 2*
 - b. Ranking B: *Correct answer: 3*
 - c. Ranking C: *Correct answer: 2*
 - d. Ranking D: *Correct answer: 2*
 - e. Ranking D: *Correct answer: 3*

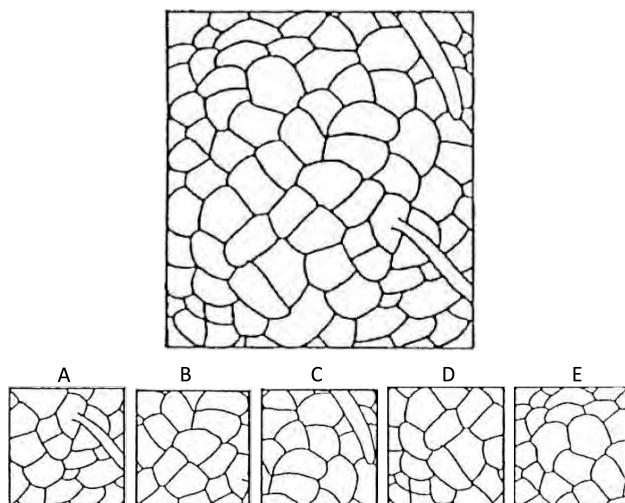
B.4 Admissions Test for Medical Studies questions

Ability Questions (1-3)

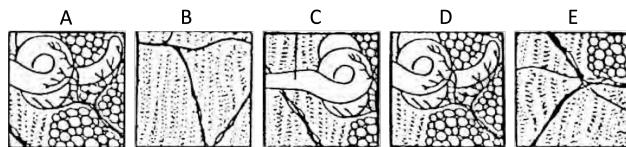
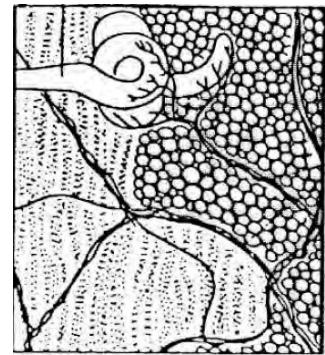
You have **three minutes**.

In the following tasks, your ability to recognize excerpts in a complex image will be tested. For each task, a “pattern” and five “pattern excerpts” (A) to (E) will be provided. You are to determine which of these five “pattern excerpts” can be placed over the pattern at any point identically and completely; the “pattern excerpts” are neither enlarged nor reduced, nor rotated or tilted.

Question 1



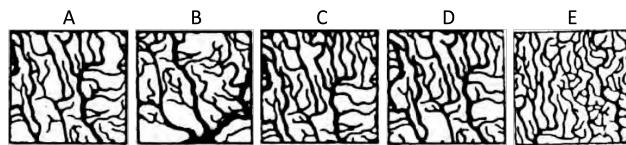
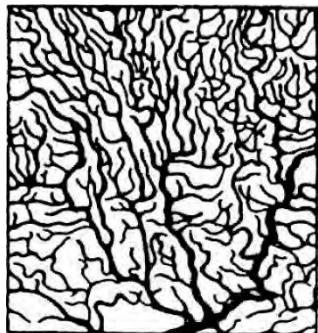
Correct answer: B



Question 2

Correct answer: A

Question 3



Correct answer: E

Ability Questions (4-6)

You have **four minutes**.

The following tasks test your ability to handle numbers, sizes, units, and formulas correctly within the context of medical and scientific questions. For each task on the answer sheet, mark the answer that is correct in the sense of the question.

Question 4

Stimuli that act on the skin from the outside are converted into bioelectric impulses in special sensory organs of the skin. The impulses generated in this way run over the afferent (incoming) nerve fibers and the so-called dorsal roots of the spinal cord into the spinal cord, where they are switched to other nerve cells. They can now trigger reflex movements via motor nerve cells; however, they can also reach the cerebral cortex after multiple switches via ascending pathways, where they are further processed and enable a conscious perception or recognition of the stimuli.

In a patient, the dorsal roots of the spinal cord are severed. Which of the following failures can be expected according to this information?

1. No bioelectric impulses are formed in the sensory organs of the skin anymore.
2. Reflex movements can no longer be triggered by stimulation.
3. Skin stimuli can no longer be consciously perceived or recognized.
 - a.) Only failure 1 is to be expected.
 - b.) Only failure 2 is to be expected.
 - c.) Only failure 3 is to be expected.
 - d.) Only failures 1 and 3 are to be expected.
 - e.) Only failures 2 and 3 are to be expected.

Correct answer: e.) Only failures 2 and 3 are to be expected.

Question 5

Among the hormones that play a significant role in regulating electrolyte and water balance is aldosterone, which is produced in the adrenal cortex and promotes the active transport of sodium ions through cell membranes. Aldosterone causes the reabsorption of sodium ions from the so-called primary urine back into the blood (the primary urine is filtered out of the blood by the kidneys). It thus reduces the excretion of sodium in the urine and sweat. An increase in aldosterone secretion is, among other things, caused by a negative sodium balance (more sodium is excreted than absorbed).

Which of the following statements can be inferred from this information?

1. The salt content (sodium chloride content) of sweat increases in the case of aldosterone deficiency.
2. A diet high in salt (sodium chloride) generally leads to increased aldosterone secretion.

3. A strong sweat secretion occurring under heat stress generally leads to decreased aldosterone production.

- a.) Only failure 1 is to be expected.
- b.) Only failure 2 is to be expected.
- c.) Only failures 1 and 3 are to be expected.
- d.) Only failures 1 and 3 are to be expected.
- e.) Only failure 3 is to be expected.

Correct answer: a.) Only failure 1 is to be expected.

Question 6

The capillaries are not only part of the blood transport system, but are also where exchange processes between blood and tissue through the vessel wall take place. At the beginning of the capillaries, there is a hydrostatic pressure difference of 30 mmHg (33 mmHg in the blood versus 3 mmHg in the tissue fluid). This outward-directed pressure is countered by the so-called “colloid osmotic pressure”, which is directed inward. It is constantly 22 mmHg across the entire capillaries. Thus, at the beginning of the capillaries, blood fluid exits the capillaries into the tissue with a resulting pressure of 8 mmHg (effective filtration pressure); at the end of the capillaries, on the other hand, a return flow of fluid from the tissue into the blood occurs under the resulting pressure of 7 mmHg (reabsorption pressure). In the case of protein deficiency malnutrition, the colloid osmotic pressure in the blood drops.

What consequences does this have for the exchange processes between capillaries and tissue?

- a.) Less fluid flows from the capillaries into the tissue, as the effective filtration pressure is lower.
- b.) More fluid transfers into the tissue, as the effective filtration pressure is higher.
- c.) The return flow of fluid into the blood is increased, since the effective filtration pressure is higher.
- d.) The return flow of fluid into the blood is decreased, since the reabsorption pressure is higher.
- e.) There is no shift in the fluid balance, as the colloid osmotic pressure along the capillaries is constant.

Correct answer: b.) More fluid transfers into the tissue, as the effective filtration pressure is higher.

Ability Questions (7-9)

You have **four minutes**.

Question 7

The term “plasma half-life” refers to the period during which the amount of a drug present in the blood plasma reduces by half; this can occur through either excretion or biological degradation. A patient is intravenously injected with a drug that has a plasma half-life of 8 hours at time t_0 . After 24 hours, there are still 10 mg of the drug in the patient’s blood plasma.

How many mg were injected into the patient?

- a.) 40 mg.
- b.) 80 mg.
- c.) 160 mg.
- d.) 200 mg.
- e.) 400 mg.

Correct answer: b.) 80 mg.

Question 8

When a direct current flows through a dilute solution of copper sulfate, metallic copper is formed at the negative pole. The amount of copper deposited is directly proportional to both the duration of the current flow and the current strength. At a current strength of 0.4 Amperes, 0.12 g of copper are deposited in 15 minutes.

How long does it take for 0.24 g of copper to be deposited at a current strength of 1 Ampere?

- a.) 6 minutes.
- b.) 12 minutes.
- c.) 20 minutes.
- d.) 30 minutes.
- e.) 75 minutes.

Correct answer: b.) 12 minutes.

Question 9

The total focal length f_g of two lenses with focal lengths f_1 and f_2 , which are at a distance d from each other, is calculated according to the formula $f_g = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 \cdot f_2}$.

If a focal length or the total focal length is positive, it indicates a converging lens or lens system, respectively; if it is negative, a diverging lens or lens system. Which statement is correct?

1. Combining two converging lenses at a distance of $d > f_1 + f_2$ results in a diverging lens system.
 2. If $f_1 = -f_2$ and $d > 0$, then $f_g = 0$.
 3. If $f_1 = f_2$ and $d > 0$, then $f_g = 2 \cdot f_1$.
 4. By choosing an appropriate distance d , a converging lens system can be created with two diverging lenses.
 5. The larger d is, while f_1 and f_2 remain constant, the larger f_g becomes.
-
- a.) A.
 - b.) B.
 - c.) C.
 - d.) D.
 - e.) E.

Correct answer: a.) A.

C Additional analyses

C.1 Comparison to Pilot Experiment

We conducted a pilot with 116 medical students from the University of Cologne between October 2017 and July 2018. Below, we discuss the design and experimental protocol of the pilot (Section C.1.1) and present the results (Section C.1.2).

C.1.1 Pilot Design and Experimental Protocol

Table C.1: Pairwise comparisons of effort choice, by ability type (Asymmetric ranking systems)

		High-ability types			
		B	T	TM	MB
Low-ability types	B	3.21 (3.59)	44%**	70%***	46%**
	T	4.80 (3.44)	-8%	4.64 (3.43)	18%**
	TM	-5%	4.40 (3.53)	3%	5.46 (3.03)
	MB	2%	11%	4.50 (3.68)	4.68 (3.09)

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} , i.e., $\bar{e}_l(R_{\text{row}})/\bar{e}_l(R_{\text{col}}) - 1$. 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).

The experiment adopted the same general design and decision situation with outcomes according to the MAIN treatment; see Section 3.2. While the incentives for the subjects were identical, the chosen effort in the pilot, denoted as $e_i^{\text{Pilot}} \in [0, 20]$, entailed linear costs $c(e_i^{\text{Pilot}}) = e_i^{\text{Pilot}}$. Thus, half of the effort chosen in the pilot has the same effect on the profit of patients and subjects as the effort in our current experiment, i.e., $e = 0.5 \cdot e^{\text{pilot}}$. To simplify comparisons, all effort choices from the pilot are therefore transformed to match the effort scale used in the main part of the paper.

We consider all ranking systems from Section 3.3. Additionally, we examine the two remaining unique combinations of thresholds to construct ranking designs: The **B ranking system** has only a single threshold at the bottom tail of the distribution, such that physicians achieving outcomes H_h, H_l , and L_h are assigned to the first rank, while only physicians with the lowest outcome L_l

are ranked last. Thus, in this ranking system, high-ability physicians always reach the first rank regardless of effort, whereas low-ability physicians can reach both the first and second ranks. The **TB ranking system** ensures high-ability physicians achieving H_h are ranked first, and low-ability physicians achieving L_l are ranked third. Physicians achieving H_l or L_h are pooled in the second rank. This ranking implies that high-ability and low-ability physicians can increase their rank through effort.

To avoid overwhelming subjects with seven different ranking designs at once, we considered two treatments. The asymmetric ranking-systems treatment, with 56 subjects, included all ranking systems where the number of achievable thresholds for the two ability types differed. These are the ranking systems B, T, TM , and MB . Conversely, the symmetric ranking-system treatment, with 60 subjects, included all rankings where the number of achievable thresholds was identical for the two ability types. These are the ranking systems M, TB , and TMB . Thus, in the pilot, the within-subject comparisons were only possible for three and four ranking systems, respectively.

Recruitment and experimental protocols were similar to those described for students in Section 3.4. However, medical students were not asked to bring their own laptops; instead, they used tablet computers from the mobile lab of the Cologne Laboratory for Experimental Research (CLER).

C.1.2 Pilot Results

Table C.2: Pairwise comparisons of effort choice, by ability type (Symmetric ranking systems)

		High-ability types		
		M	TB	TMB
Low-ability types	M	4.60 (2.30)	6%*	23%**
	TB	4.10 (2.70)	3%	4.74 (2.10)
	TMB	9%	13%	5.04 (2.20)

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} , i.e., $\bar{e}_l(R_{\text{row}})/\bar{e}_l(R_{\text{col}}) - 1$. 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).

Table C.1 compares the asymmetric ranking systems. Similar to Result 1, we find that adding

achievable thresholds (tends to) increase effort for both low-ability subjects ($\bar{e}_l(B) < \bar{e}_l(MB)$ and $\bar{e}_l(T) < \bar{e}_l(TM)$) as well as high-ability subjects ($\bar{e}_h(B) < \bar{e}_h(MB)$ and $\bar{e}_h(T) < \bar{e}_h(TM)$). However, these comparisons are not always statistically significant. We also observe that having no achievable threshold for low-ability subjects, i.e., the T ranking system, leads to the lowest effort. Similarly, having no threshold that high-ability subjects can fall below, i.e., the B ranking system, results in the lowest effort for high-ability subjects emphasizing the importance of including at least one relevant threshold for each type to increase effort levels.

Table C.2 compares the symmetric ranking systems. Similar to Result 1, we again find that, for both ability types, additional thresholds motivate as $\bar{e}_l(TB) < \bar{e}_l(TMB)$ and $\bar{e}_h(TB) < \bar{e}_h(TMB)$. Moreover, as with Result 2, additional relevant thresholds tend to motivate more than additional unachievable thresholds, as $\bar{e}_l(M) < \bar{e}_l(TMB)$ and $\bar{e}_h(M) < \bar{e}_h(TMB)$. For high-ability subjects, even significantly so.

Finally, Table C.3 describes the effect of introducing rankings relative to the non-ranking baseline. We observe more positive effects than in the main part of the paper. Particularly, no ranking system significantly decreases effort relative to the baseline. Examining the two types separately, we find that high-ability subjects increase effort relative to the baseline if an achievable threshold is included – albeit, not always significantly. On the other hand, low-ability physicians tend to react positively to a ranking if the bottom threshold is included, even significantly so for the TMB ranking system.)

Table C.3: Effort choices under rank feedback versus the non-ranking baseline

	All abilities		High-ability		Low-ability	
	Mean (s.d.)	%-Diff to Base	Mean (s.d.)	%-Diff to Base	Mean (s.d.)	%-Diff to Base
A. Asymmetric rankings ($N = 56$)						
<i>Non-ranking baseline</i>						
Base	4.14 (3.25)		3.75 (3.05)		4.54 (3.44)	
Ranking						
B	4.00 (3.57)	-3.38	3.21 (3.59)	-14.72	4.80 (3.44)	5.73
T	4.52 (3.45)	9.18	4.64 (3.43)	23.87**	4.40 (3.53)	-3.08
TM	5.00 (3.37)	20.77***	5.46 (3.03)	45.73***	4.50 (3.68)	-0.88
MB	4.78 (3.26)	15.46***	4.68 (3.09)	24.80**	4.90 (3.47)	7.93
B. Symmetric rankings ($N = 60$)						
<i>Non-ranking baseline</i>						
Base	4.10 (3.36)		4.07 (2.00)		4.14 (2.20)	
Ranking						
M	4.35 (3.46)	6.10	4.60 (2.30)	13.02	4.10 (2.70)	-0.97
TB	4.55 (3.57)	10.98*	4.74 (2.10)	16.46	4.36 (2.50)	5.31
TM/B	5.07 (3.61)	23.54***	5.04 (2.20)	23.83***	5.10 (2.70)	23.19**

Notes. This table shows descriptive statistics for physicians' baseline efforts (without a ranking) and efforts under the ranking designs for the two ranking designs. In the baseline decisions, subjects made two decisions (before the availability was determined in a real-effort task), as if they were a low-ability and a high-ability physician. Here, we show descriptive statistics and non-parametric test results for the subjects' efforts according to their type determined in a real-effort task. The relative changes in effort compared to baseline are given in percentages. p -values are shown for two-sided Wilcoxon signed-rank test with Holm-correction.

C.2 Comparison to non-ranking baseline

Table C.4: The effect of ranking designs on physician effort relative to the baseline

Model	All abilities	High-ability	Low-ability
	(1)	(2)	(3)
<i>T</i>	-0.969*** (0.166)	0.460*** (0.177)	-2.398*** (0.236)
<i>M</i>	-0.202* (0.119)	0.102 (0.135)	-0.506*** (0.195)
<i>TM</i>	0.006 (0.129)	0.682*** (0.156)	-0.670*** (0.193)
<i>MB</i>	-0.014 (0.124)	0.057 (0.143)	-0.085 (0.205)
<i>TMB</i>	0.341*** (0.130)	0.864*** (0.158)	-0.182 (0.200)
Altruism	0.476*** (0.091)	0.267** (0.128)	0.707*** (0.128)
Female	0.311 (0.281)	0.417 (0.392)	0.175 (0.382)
Constant	4.940*** (0.310)	5.058*** (0.411)	4.767*** (0.449)
Observed decisions	2112	1056	1056
Subjects	352	176	176

Notes. This table shows estimation results from Tobit regressions with robust standard errors clustered at the individual subject level. The reference category is *T*. We include a gender dummy that equals 1 for females and 0 for males. We also include a measure for altruism which were elicited using incentivized dictator game after the main experiment. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

C.3 Robustness of results to other parametric specifications

This subsection includes full regression of the main part and estimating of the models in the main part using Tobit rather than OLS.

Table C.5: Tobit regression on the effect of ranking designs on effort relative to the T-ranking

Model	All abilities			High-ability			Low-ability		
	Phys. (1)	Stud. (2)	Control (3)	Phys. (4)	Stud. (5)	Control (6)	Phys. (7)	Stud. (8)	Control (9)
<i>M</i>	0.526** (0.281)	1.075*** (0.109)	1.247*** (0.146)	-0.077 (0.099)	-0.764*** (0.092)	-0.484*** (0.099)	1.154*** (0.281)	3.029*** (0.278)	3.155*** (0.441)
<i>TM</i>	0.746*** (0.200)	1.219*** (0.070)	1.671*** (0.109)	0.107* (0.055)	-0.045 (0.029)	0.634*** (0.047)	1.411*** (0.200)	2.558*** (0.224)	2.836*** (0.405)
<i>MB</i>	0.882*** (0.254)	1.162*** (0.143)	1.520*** (0.162)	0.053 (0.090)	-1.143*** (0.134)	-0.334*** (0.112)	1.759*** (0.254)	3.604*** (0.320)	3.566*** (0.432)
<i>TM*MB</i>	1.053*** (0.246)	1.723*** (0.095)	2.107*** (0.179)	0.695*** (0.084)	0.076* (0.032)	0.734*** (0.086)	1.420*** (0.246)	3.449*** (0.294)	3.611*** (0.662)
Altruism	0.742*** (0.120)	0.532*** (0.038)	0.447*** (0.046)	0.717*** (0.103)	0.234*** (0.052)	-0.167*** (0.076)	0.884*** (0.120)	0.771*** (0.087)	1.260*** (0.052)
Female	1.498 (0.963)	0.327 (0.268)	-0.596 (0.422)	1.876* (0.763)	-0.342 (0.341)	-0.166 (0.652)	1.084 (0.963)	0.920 (0.573)	-1.916* (1.107)
Constant	2.954 (1.819)	3.981*** (0.535)	3.852*** (0.387)	3.791* (1.561)	7.162*** (0.546)	5.655*** (0.547)	1.872 (1.819)	0.872 (1.076)	2.043*** (0.773)
<i>Differences between coefficients</i>									
<i>Wald tests of the following hypotheses H_0:</i>									
$M = TM$	-0.221	-0.144	-0.424**	-0.184	-0.720***	-1.118***	-0.258	0.470*	0.318
$M = MB$	-0.356*	-0.087	-0.273*	-0.131	0.378*	-0.150	-0.605	-0.575**	-0.412*
$M = TMB$	-0.527*	-0.648***	-0.849***	-0.772*	-0.840***	-1.218***	-0.266	-0.420	-0.456
$TM = MB$	-0.135	0.057	0.151	0.054	1.038***	0.968***	-0.348	-1.045***	-0.730*
$TM = TMB$	-0.307	-0.504***	-0.436	-0.588*	-0.120	-0.100	-0.008	-0.891**	-0.774
$MB = TMB$	-0.171	-0.561***	-0.587***	-0.642***	-1.218***	-1.068***	0.340	0.155	-0.045
Observed decisions	560	640	560	275	320	280	280	320	280
Subjects	112	128	112	56	64	56	56	64	56

Notes. This table shows estimation results from Tobit regressions with robust standard errors clustered at the individual subject level. The reference category is T. We include a gender dummy that equals 1 for females and 0 for males. We also include a measure for altruism which were elicited using incentivized dictator game after the main experiment. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.