

# Kölner Impulse zur Wirtschaftspolitik

Nr. 6/2024 | 3. Dezember 2024

## In diesem Impuls...

... beleuchten Yero Ndiaye, Katharina Huesmann, Christian Waibel und Daniel Wiesen den Einfluss von Feedback und Rankingsystemen auf die Anstrengung von Ärztinnen und Ärzten. Nach einer Einführung in die Wirkung von Leistungsfeedback und die Mechanismen sozialer Vergleiche im klinischen Kontext analysieren sie, wie sich unterschiedliche Rankingsysteme auf die Motivation von Ärztinnen und Ärzten auswirken. Ihre Ergebnisse unterstreichen die Bedeutung einer sorgfältigen Anpassung von Feedbacksystemen an die Fähigkeiten des Teams, um Motivation und die Behandlungsqualität in Krankenhäusern nachhaltig zu steigern.

Außerdem kommentieren unsere Direktoren, Prof. Michael Krause Ph.D. und Prof. Dr. Steffen Roth, kurz das Gutachten des Sachverständigenrats.



## Neues aus dem iwp

Am 9. Januar 2025 findet die Ludwig Erhard Lecture 2025 an der Universität zu Köln statt. Um 17:45 Uhr in Aula II im Hauptgebäude der UzK wird uns Prof. Dr. Dr. h.c. Clemens Fuest, Präsident des ifo Instituts für Wirtschaftsforschung, spannende Einblicke in das Thema „Wirtschaftsaussichten 2025 und der Wirtschaftsstandort Deutschland“ gewähren. Die Anmeldung zur Veranstaltung sowie weitere Informationen finden sich [hier](#).

Im Dezember ist soeben Heft 3/2024 der Zeitschrift für Wirtschaftspolitik erschienen u.a. mit dem aktuellen Forumsthema "STREIK – ein gutes Recht?" mit Beiträgen von Yasmin Fahimi, Erwin Amann und Hagen Lesch.

Das iwp wünscht allen Lesenden eine schöne und besinnliche Weihnachtszeit!

# Kurz kommentiert: Zum Gutachten des Sachverständigenrats

Michael Krause und Steffen J. Roth

Vor kurzem hat der Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung sein Jahresgutachten vorgestellt, unter dem Motto: Versäumnisse angehen, entschlossen modernisieren. Klar ist, dass die stagnierende deutsche Wirtschaft vor konjunkturellen und vor allem strukturellen Problemen steht. Dabei handelt es sich im Wesentlichen nicht um eine UnterAuslastung von Produktionskapazitäten, sondern um einen einsetzenden Rückgang des gesamtwirtschaftlichen Produktionspotentials. Daher sehen wir auch noch keinen nennenswerten Anstieg der Arbeitslosigkeit. Der Mangel an Fachkräften ist ein Grund für den Rückgang der Kapazitäten. Auch bleibt ein bei konjunkturellen Abschwüngen zu erwartender Rückgang der Inflation aus. Da keine herkömmliche Rezession vorliegt, wird ihr auch nicht mit einer Abfederung mittels traditioneller Konjunkturpolitik beizukommen sein.

Wie das Gutachten betont, müssen zukunftsorientierte öffentliche Ausgaben priorisiert werden, die insbesondere die Bereiche Infrastruktur, Bildung und Verteidigung betreffen. Von den ersten beiden können Wirkungen auf das zukünftige Produktionspotential erwartet werden. Die Verteidigungsfähigkeit wird helfen, zukünftige Konflikte zu vermeiden.

Das Problem zukunftsorientierter öffentlicher Ausgaben ist eben jene Ausrichtung auf die Zukunft: Kosten entstehen kurzfristig, der Nutzen wird bestenfalls mittelfristig sichtbar. Aus dem gleichen Grund konnte die Infrastruktur so lange vernachlässigt werden: Die Erneuerung einer Brücke oder Autobahn mag sechs- oder siebenstellige Beträge verschlingen, während ein Aufschieben dieser Maßnahmen zunächst kaum merklich ist.

Für die Politik sind zukunftsorientierte öffentliche Ausgaben denkbar unattraktiv, da die langfristigen Erfolge die kurzfristigen Anstrengungen vor der nächsten Wahl nicht sichtbar ausgleichen. Wie aber sollte die Politik glaubwürdig Versprechen machen können, in der nächsten Legislaturperiode langfristig sinnvolle Maßnahmen zu ergreifen?

Aus unserer Sicht sind im Gutachten des Sachverständigenrats neben vielen wichtigen konkreten Vorschlägen auch Maßnahmen aufgeführt, die mittelfristig aus dem Dilemma führen können, dass mit den eigentlich wichtigen Investitionen keine Wahlen zu gewinnen zu sein scheinen. Diese Maßnahmen kosten beim Wähler nichts, führen aber – hoffentlich – zu besseren Entscheidungen.

Es geht uns um die im Gutachten formulierten Forderungen nach Mechanismen zu Ex-ante Evaluierung mithilfe von Kosten-Nutzen-Analysen, mit denen die Priorisierung von Ausgaben erleichtert werden soll. Und mehr noch nach einer kontinuierlichen Ex-post Prüfung und Evaluierung des Mitteleinsatzes. Es ist auf diese Weise zu hoffen, dass die Regierung und das Parlament im Bewusstsein zukünftiger Evaluationen bessere Entscheidungen treffen, da nicht mehr davon ausgegangen werden kann, dass sie aus dem Bewusstsein der Bevölkerung schwinden.

Ein ergänzendes Element sind hierbei die vorgeschlagenen Auslaufklauseln (Sunset Clauses) für Gesetze, die die Regierenden nötigen, die Evaluation ernst zu nehmen, und ein Weiterführen oder Verändern von fällig gewordenen Gesetzen im Parlament und der Öffentlichkeit rechtfertigen zu müssen. Eine solche Selbstbindung schafft die erwünschte Glaubwürdigkeit und sollte eine disziplinierende Wirkung entfalten. Ob zumindest eine der Parteien diese Pläne im anstehenden Wahlkampf auf ihre Fahnen schreiben wird, bleibt abzuwarten.

# Wann kann Feedback zur Leistung Ärztinnen und Ärzte motivieren?

---

Von Yero S. Ndiaye, Katharina Huesmann, Christian Waibel und Daniel Wiesen

## Wie kann man Leistungen von Ärztinnen und Ärzten verbessern?

In Krankenhäusern ist die hohe Qualität der Patientenversorgung das wichtigste Ziel. Eine zentrale Rolle bei der Erreichung dieses Zieles nehmen die Ärztinnen und Ärzte ein, deren Fähigkeiten und Anstrengung die Qualität der Patientenversorgung maßgeblich bestimmen. Eine der Managementmaßnahmen, die Krankenhausleitungen zur Steigerung der Qualität der Versorgung zur Verfügung stehen, ist Feedback in Form von relativen Rankings.

Die zugrundeliegende Idee ist einfach: Feedback durch Rankings, das die eigene Leistung im Vergleich zu anderen innerhalb einer Gruppe oder eines Teams transparent macht, soll den sozialen Vergleich fördern. Dieser soll wiederum die Motivation steigern, sich zu verbessern. Wer beispielsweise erfährt, dass er oder sie im Vergleich zu anderen hinterherhinkt, könnte versuchen, durch noch höhere Anstrengung aufzuholen. Allerdings kann dieser Ansatz auch unerwünschte demotivierende Reaktionen hervorrufen.

In einem Ranking wird jedem Arzt und jeder Ärztin entsprechend der erreichten Leistung in einem individuell messbaren Qualitätsindikator ein Rang zugeordnet. Die Ranggrenzen legen fest, welche ärztliche Leistung bei einer messbaren Kennzahl zu welchem Rang führt. Die Gestaltung eines solchen Rankings stellt eine Herausforderung für die Führungskräfte dar: Eine zusätzliche hohe Ranggrenze kann motivierend auf die Ärztinnen und Ärzte wirken, die eine realistische Chance haben, diese Grenze zu überschreiten. Andererseits kann sie diejenigen demotivieren,

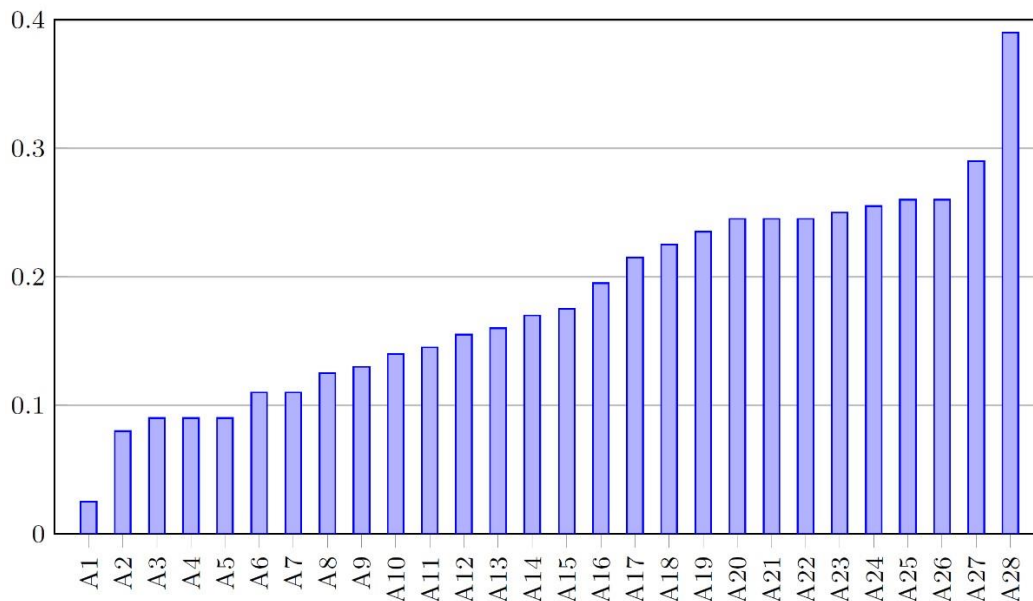
die glauben, das Ziel selbst bei hoher Anstrengung nicht erreichen zu können.

Es entsteht ein Dilemma: Bestimmte Rankings könnten zwar einige Teammitglieder zu besseren Leistungen anspornen, andere jedoch demotivieren. Unterschiedliche Fähigkeiten im Team machen dieses Problem noch komplexer, da Ärztinnen und Ärzte aufgrund von Talent oder Erfahrung trotz gleicher Anstrengung oft unterschiedliche medizinische Behandlungsergebnisse erzielen.

Obwohl medizinische Fachgesellschaften die Einführung von Peer Feedback durch Rankings häufig empfehlen (z.B. Royal College of Surgeons in Großbritannien), steckt die Implementierung immer noch in den Kinderschuhen. Ein Beispiel eines klinischen Bereichs, in dem Rankings eingeführt werden könnten, bietet die Gastroenterologie; siehe Abbildung 1. Die Adenom-Entdeckungsrate – also der Anteil an Darmspiegelungen, bei denen Adenome entdeckt werden – gilt als wichtiger Qualitätsindikator. Nach britischen Richtlinien liegt ein akzeptabler Bereich zwischen 10 % und 20 % und sehr gute Leistungen über 25 %. Würde das Ranking jedoch nur eine Ranggrenze von 25 % einführen, könnten nur wenige den ersten Rang erreichen, was die meisten Kolleginnen und Kollegen eher entmutigt. Nur eine Ranggrenze bei 10 % hingegen wäre für die Mehrheit leicht erreichbar und würde kaum Anreiz bieten, sich anzustrengen.

Ein gut durchdachtes Ranking sollte daher sorgfältig auf das Leistungsspektrum des Teams abgestimmt sein. Die Herausforderung liegt darin, die richtige Balance zu finden, um möglichst viele

**Abbildung 1: Entdeckungsrate von Adenomen in einem britischen Krankenhaus**



Entdeckungsrate der Adenome von 28 Ärztinnen und Ärzten (A1 bis A28) in einem britischen Krankenhaus

Ärztinnen und Ärzte zu motivieren, ohne einen Teil der Belegschaft zu frustrieren.

In unserer Studie (Peer Feedback in Healthcare: How Ranking Designs and Ability Affect Physician Effort<sup>1</sup>) analysieren wir in einem Experiment mit 112 praktizierenden Ärztinnen und Ärzten sowie 240 zukünftigen Ärztinnen und Ärzten (Medizinstudentinnen und Medizinstudenten), wie sich Rankings und Fähigkeiten auf Anstrengung auswirken. Aus unseren Resultaten folgen wichtige Empfehlung für Führungskräfte, die Feedback in Form von Rankings in ihren Krankenhäusern einführen wollen.

## Was über Feedback (noch nicht) bekannt ist

In der Forschung zur Leistungssteigerung im Gesundheitswesen und der Verhaltensökonomie gibt es viele Studien zu Feedbacksystemen, die medizinische Fachkräfte über ihre Leistung im Vergleich zu Kolleginnen und Kollegen informieren – dabei aber auf finanzielle Anreize für gute

Leistungen verzichten. Einige Studien zeigen, dass dieses „relative Leistungsfeedback“ die Leistung in Krankenhäusern und medizinischen Einrichtungen positiv beeinflussen kann. Beispielsweise fand eine Studie<sup>2</sup> heraus, dass die öffentliche Bekanntgabe der Verweildauer von Notfallpatientinnen und -patienten in Verbindung mit dem Austausch von Best-Practice-Beispielen die Produktivität in Notaufnahmen steigert. Auch im Bereich der Primärversorgung und bei Grippeimpfungen führte solches Feedback zu besseren Ergebnissen als finanzielle Anreize.<sup>3</sup>

In experimentellen Studien wurde gezeigt, dass öffentliches Feedback oft die Leistung verbessert, weil es sozialen Vergleich unter Kolleginnen verstärkt – niemand will schlecht dastehen<sup>4</sup>. Privat gegebenes Feedback, d.h. ohne Information für die anderen Teammitglieder, kann durch den Einfluss auf das Selbstbild ebenfalls leistungssteigernd wirken<sup>5</sup>.

<sup>1</sup> Huesmann et al. (forthcoming)

<sup>2</sup> Song et al. (2018)

<sup>3</sup> Navathe et al. (2020), Niewoehner und Staats (2022)

<sup>4</sup> Hannan et al. (2013), Tafkov (2013), Gerhards und Siemer (2016)

<sup>5</sup> Tafkov (2013), Gill et al. (2019), Kuhnén und Tymula (2012)

Auf der anderen Seite zeigen aber auch viele Studien, dass Leistungsfeedback gar keinen oder sogar einen negativen Effekt auf die Leistung hat<sup>6</sup>. Eine Untersuchung im klinischen Bereich ergab zum Beispiel, dass Ärztinnen und Ärzte besser auf positives Feedback reagieren, während negatives Feedback zu schlechteren Ergebnissen führen kann<sup>7</sup>.

Ein häufig untersuchter Aspekt ist die Reaktion auf das Erreichen von Spitzenpositionen oder das Vermeiden von Schlusspositionen: Besonders gute oder schlechte Platzierungen führen in der Folge oft zu großen Leistungssteigerungen. Intuitiv kann der Ehrgeiz, an allererster Stelle zu stehen, sowie die Angst, auf der allerletzten Stelle zu landen, zu einer besonders hohen Anstrengung führen<sup>8</sup>.

Insgesamt deuten diese widersprüchlichen Befunde darauf hin, dass Feedback unterschiedlich wirken kann. Ein Grund dafür könnte das Design des Feedbacksystems sein, dem bisher wenig Aufmerksamkeit gewidmet wurde. Üblicherweise werden nämlich lediglich zwei Arten von Ranking-Systemen untersucht: entweder vollständig granulare Systeme (in denen jedes mögliche Behandlungsergebnis eine eigene Rangstufe erhält) oder solche, die nur die Top-Performer auszeichnen.

## Wie Statusmotive die Anstrengung beeinflussen können

Unsere Studie beantwortet nun die bisher noch nicht untersuchte Frage, wie Ärztinnen und Ärzte mit unterschiedlicher Leistungsfähigkeit auf Feedback in unterschiedlichen Rankings mit einer verschiedenen Anzahl von Ranggrenzen reagieren.

Ein Ranking funktioniert, indem bestimmte Ranggrenzen definiert werden, anhand derer die Leistung von Ärztinnen und Ärzten in verschiedene Ränge eingeteilt werden. Unsere Hypothesen

konzentrieren sich darauf, wie die Einführung einer neuen Ranggrenze die Anstrengung beeinflusst – und zwar je nach Fähigkeit der Ärztinnen und Ärzte, diesen neuen Rang zu erreichen.

Unsere erste Hypothese basiert auf der Annahme, dass höhere Ränge mehr Status im Team bedeuten, was zu mehr Anstrengung anspornt. Fügen wir also eine neue Ranggrenze hinzu, erhalten die Teammitglieder eine weitere Möglichkeit, sich durch höhere Anstrengung abzuheben. Dabei gilt jedoch: Wenn die neue Ranggrenze zu hoch ist, wird dies diejenigen demotivieren, die sie ohnehin nicht erreichen können. Unsere erste Hypothese lautet daher:

**Hypothese 1:** Das Hinzufügen einer neuen Ranggrenze in einem Ranking beeinflusst die Anstrengung von Ärztinnen und Ärzten. Die Richtung des Effekts hängt davon ab, ob eine Ärztin oder ein Arzt die neue Ranggrenze erreichen kann:

- a) Ärztinnen und Ärzte, die Leistungen ober- und unterhalb der neuen Ranggrenze erreichen können, erhöhen ihre Anstrengung.
- b) Ärztinnen und Ärzte, die die Ranggrenze nicht erreichen können, reduzieren ihre Anstrengung.
- c) Ärztinnen und Ärzte, die die Ranggrenze problemlos überschreiten können, reduzieren ebenfalls ihre Anstrengung, da weniger Anstrengung nötig ist, um den gleichen Status zu erreichen.

Da innerhalb eines Teams erhebliche Unterschiede in der Leistungsfähigkeit bestehen könnten, gibt es ein Dilemma: Ranggrenzen in der Mitte des Leistungsspektrums könnten für alle ein erreichbares, aber herausforderndes Ziel darstellen. Ranggrenzen an den Enden des Spektrums hingegen könnten nur wenige motivieren und den Rest demotivieren.

Die Attraktivität extremer Ranggrenzen könnte dieses Dilemma lösen: Die Aussicht, entweder einen besonders hohen Rang zu erreichen oder das

---

<sup>6</sup> Bandiera et al. (2013), Ashraf et al. (2014), Charness et al. (2014), Edelman and Larkin (2015), Turkoglu and Tucker (2022)

<sup>7</sup> Singh and Zureich (2023)

<sup>8</sup> Azmat and Iriberry (2010), Kuziemko et al. (2014), Gill et al. (2019), Niewoehner and Staats (2022)



untere Ende des Rankings zu vermeiden, kann besonders motivierend sein. Dies führt uns zu unserer zweiten Hypothese:

**Hypothese 2:** Wenn eine Ranggrenze in der Nähe des höchsten oder niedrigsten Behandlungsergebnisses gesetzt wird, ist der Anstieg der Anstrengung von Ärztinnen und Ärzten, die die Ranggrenze erreichen können, größer als der Rückgang der Anstrengung bei den Ärztinnen und Ärzten, die sie nicht erreichen können. Werden Ranggrenzen sowohl am oberen als auch am unteren Ende des Leistungsspektrums hinzugefügt, steigt die Anstrengung bei beiden Gruppen.

Mit diesen Hypothesen wollen wir die bestehenden Erkenntnisse aus der Literatur weiter differenzieren, indem wir untersuchen, wie das Design des Rankings das Verhalten beeinflusst. Zudem lässt unser experimentelles Design den Vergleich der Anstrengung mit und ohne Feedback in Form von Rankings zu, wodurch der reine Effekt einer Einführung von relativem Leistungsfeedback gemessen werden kann.

## Ein stilisiertes Experiment klinischer Situationen

Unser Experiment simuliert Entscheidungen, wie sie im Gesundheitswesen getroffen werden könnten, in drei Stufen. In Stufe 1 und 3 wählten die Teilnehmerinnen und Teilnehmer mehrmals Anstrengungen von 0 bis 10, um eine abstrakte Patientin zu behandeln. Höhere Anstrengung war mit höheren Kosten verbunden (jede Einheit Anstrengung kostete zwei experimentelle Einheiten (ECU)), aber sie erhöhte die Wahrscheinlichkeit, dass die Patientin ein besseres Behandlungsergebnis erreichte.

Die Behandlungsergebnisse hingen sowohl von der Anstrengung als auch von der individuellen Fähigkeit der Teilnehmerinnen und Teilnehmer ab. Es gab zwei Fähigkeitsstufen: hoch und niedrig. Personen mit hoher Fähigkeit hatten bei gleicher Anstrengung eine höhere Wahrscheinlichkeit, bessere Behandlungsergebnisse zu erzielen.

Die Behandlungsergebnisse der abstrakten Patientinnen wurden in Geldbeträge umgerechnet und an die Christoffel-Blindenmission gespendet. Diese Organisation nutzt die Mittel zur Behandlung von Grauem Star. Dieser Ansatz stärkte das Bewusstsein der Teilnehmerinnen und Teilnehmer für den Zusammenhang zwischen ihrer Anstrengung und tatsächlichen gesundheitlichen Konsequenzen für Menschen außerhalb des Labors.

Teilnehmerinnen und Teilnehmer erhielten für sich selbst 20 ECU pro behandelte Patientin minus der gewählten Kosten der Anstrengung, unabhängig vom Behandlungsergebnis der behandelten Patientin. Diese Einheiten wurden in echtes Geld umgerechnet – mit unterschiedlichen Kursen für Medizinstudierende und berufstätige Ärztinnen und Ärzte.

### Stufe 1: Baseline-Messung ohne Rankings

Zunächst wählten die Teilnehmerinnen und Teilnehmer Anstrengungsniveaus, ohne zu wissen, welcher Fähigkeitsgruppe sie angehörten. Sie trafen zwei Entscheidungen: einmal in der Annahme, dass sie hohe Fähigkeiten haben, und einmal, dass sie niedrige Fähigkeiten haben. Diese Stufe lieferte die Basis für ihr Verhalten ohne Einfluss eines Rankings.

### Stufe 2: Bestimmung der Fähigkeiten

Die Fähigkeit wurde durch einen Wissenstest zugeordnet, der neun Fragen aus dem deutschen Mediziner-Eignungstest enthielt. Innerhalb von festen Vierergruppen wurden die zwei Personen mit den meisten richtigen Antworten als "hohe Fähigkeiten" und die anderen zwei als "niedrige Fähigkeiten" eingestuft. Die Teilnehmerinnen und Teilnehmer erfuhren anschließend nur die Fähigkeitsstufe, der sie selbst zugeordnet wurden, und welche Behandlungsergebnisse sie theoretisch in der nächsten Stufe erreichen könnten.

Zusätzlich wurden die Teilnehmerinnen und Teilnehmer gebeten, ihre Vornamen laut auszusprechen und sie in den Computer einzugeben. Dies

machte ihre Identität innerhalb ihrer Vierergruppe bekannt, um den sozialen Vergleich zu ermöglichen.

### Stufe 3: Entscheidungen unter verschiedenen Rankings

In der letzten Stufe wurden den Teilnehmerinnen und Teilnehmern fünf unterschiedliche Rankings vorgestellt, die die Leistungen in Ränge unterteilten. Diese Systeme wurden in zufälliger Reihenfolge auf dem Bildschirm angezeigt. Die Teilnehmerinnen und Teilnehmer trafen ihre Anstrengungsentscheidungen, wissend, dass sie am Ende des Experiments für eine der vorherigen Rankingentscheidung das Resultat innerhalb ihrer Vierergruppe erfahren würden. Nach Abschluss der Entscheidungen wurde eines der Rankings zufällig ausgewählt, und die Behandlungsergebnisse wurden öffentlich innerhalb der Vierergruppe bekanntgegeben.

### Die Rankings im Experiment

In unserem Experiment gibt es vier mögliche Behandlungsergebnisse, die von der Fähigkeit und Anstrengung der Teilnehmerinnen und Teilnehmer abhängen. Teilnehmerinnen und Teilnehmer mit hohen Fähigkeiten können ein sehr hohes (25 ECU) oder niedriges Behandlungsergebnis (5 ECU) für ihre Patientin erreichen. Teilnehmerinnen und Teilnehmer mit niedrigen Fähigkeiten können ein hohes (20 ECU) oder ein sehr niedriges Behandlungsergebnis (0 ECU) für ihre Patientin erreichen. Für beide Fähigkeitstypen steigt die Wahrscheinlichkeit, das (sehr) hohe Behandlungsergebnis zu erreichen, mit der gewählten Anstrengung.

Wir nutzen Ranggrenzen, um die Behandlungsergebnisse in verschiedene Ränge zu unterteilen:

- Obere Ranggrenze (T): trennt das sehr hohe vom hohen Behandlungsergebnis.
- Mittlere Ranggrenze (M): trennt das hohe und niedrige Behandlungsergebnis.
- Untere Ranggrenze (B): trennt das niedrige und sehr niedrige Behandlungsergebnis.

Aus Kombinationen der verschiedenen Ranggrenzen leiten wir fünf Rankings her; siehe Abbildung 2:

1. T-Ranking (Obere Ranggrenze)  
Nur Teilnehmerinnen und Teilnehmer mit sehr hohen Behandlungsergebnissen erreichen den ersten Rang. Alle anderen werden unabhängig von ihrer Anstrengung in den zweiten Rang eingestuft. Hier können nur Teilnehmerinnen und Teilnehmer mit hohen Fähigkeiten den ersten Rang erreichen.
2. M-Ranking (Mittlere Ranggrenze)  
Teilnehmerinnen und Teilnehmer mit hohen oder sehr hohen Behandlungsergebnissen kommen in den ersten Rang, während diejenigen mit niedrigen oder sehr niedrigen Behandlungsergebnissen im zweiten Rang landen. Hier können beide Fähigkeitsstufen in den ersten Rang aufsteigen.
3. TM-Ranking (Obere und mittlere Ranggrenze)  
Sehr hohe Behandlungsergebnisse führen zum ersten Rang, hohe Behandlungsergebnisse zum zweiten Rang, und niedrige sowie sehr niedrige Behandlungsergebnisse landen im dritten Rang. Dadurch können Teilnehmerinnen und Teilnehmer beider Fähigkeitsstufen ihre Position verbessern, jedoch nur Teilnehmerinnen und Teilnehmer mit hohen Fähigkeiten den ersten Rang erreichen.

**Abbildung 2: Rankings in unserem Experiment**

T-Ranking	M-Ranking	TM-Ranking	MB-Ranking	TMB-Ranking
Rang 1	Rang 1	Rang 1	Rang 1	Rang 1
25	25	25	25	25
Rang 2		Rang 2		Rang 2
20	20	20	20	20
5	Rang 2	Rang 3	Rang 2	Rang 3
0	5	15	5	5
	0	10	0	0
			Rang 3	Rang 4
				0

4. MB-Ranking (Mittlere und untere Ranggrenze)  
Hohe und sehr hohe Behandlungsergebnisse werden im ersten Rang eingestuft, niedrige im zweiten und sehr niedrige im dritten. Hier können Teilnehmerinnen und Teilnehmer beider Fähigkeitsstufen den ersten Rang erreichen.
5. TMB-Ranking (Alle Ranggrenzen)  
Jedes Behandlungsergebnis hat einen eigenen Rang: sehr hohe Behandlungsergebnisse (1.), hohe (2.), niedrige (3.) und sehr niedrige (4.). Dieses System ist granular und gibt vollständige Rückmeldung über das Behandlungsergebnis.

## Verschiedene Rankings und verschiedene Fähigkeiten führen zu unterschiedlichen Behandlungsergebnissen

Die Behandlungsergebnisse des Experiments werden in Abbildung 3 zusammengefasst. Zunächst zeigt sich, dass Rankings die Anstrengung von Ärztinnen und Ärzten beeinflussen, und sowohl motivierend wie auch demotivierend wirken können. Wie sich die Anstrengung verändert, hängt sowohl vom Ranking als auch von den Fähigkeiten der Ärztinnen und Ärzte ab.

Ärztinnen und Ärzte mit hohen Fähigkeiten, siehe Panel B von Abbildung 3, zeigen die größte Anstrengung, wenn Rankings die obere Ranggrenze (T) für den Spitzenplatz enthalten, den nur sie erreichen können. Die mittlere Ranggrenze (M) hinzuzufügen erhöht die Anstrengung – allerdings nur leicht (3%) und insignifikant. Die untere Ranggrenze (B) hat gemischte – aber nicht eindeutig negative Effekte. Folglich führt das granulare

Ranking (TMB-Ranking) mit allen Ranggrenzen zu der signifikant höchsten Anstrengung.

Für Ärztinnen und Ärzte mit niedrigen Fähigkeiten, siehe Panel C von Abbildung 3, ist die Anzahl erreichbarer Ranggrenzen entscheidend. Rankings, die nur das für sie unerreichbare Spitzenergebnis belohnen, d.h. nur die obere Ranggrenze (T) implementieren, führen zu einer drastischen Reduktion der Anstrengung: zwischen 41 % und 55 % weniger als bei Rankings mit zusätzlich entweder der mittleren (M) oder der unteren (B) Ranggrenze. Das Hinzufügen einer unteren Ranggrenze (z. B. von M zu MB) steigert die Anstrengung um 7–8 %, was auf eine Aversion gegen den letzten Platz hinweist. Allerdings demotiviert die hohe Ranggrenze Ärztinnen und Ärzte mit niedrigen Fähigkeiten nur leicht und insignifikant, wenn das Ranking gleichzeitig für sie erreichbare Ranggrenzen enthält. Somit führt auch hier das granulare Ranking (TMB-Ranking) mit allen Ranggrenzen zu den zwei höchsten Anstrengungsniveaus.

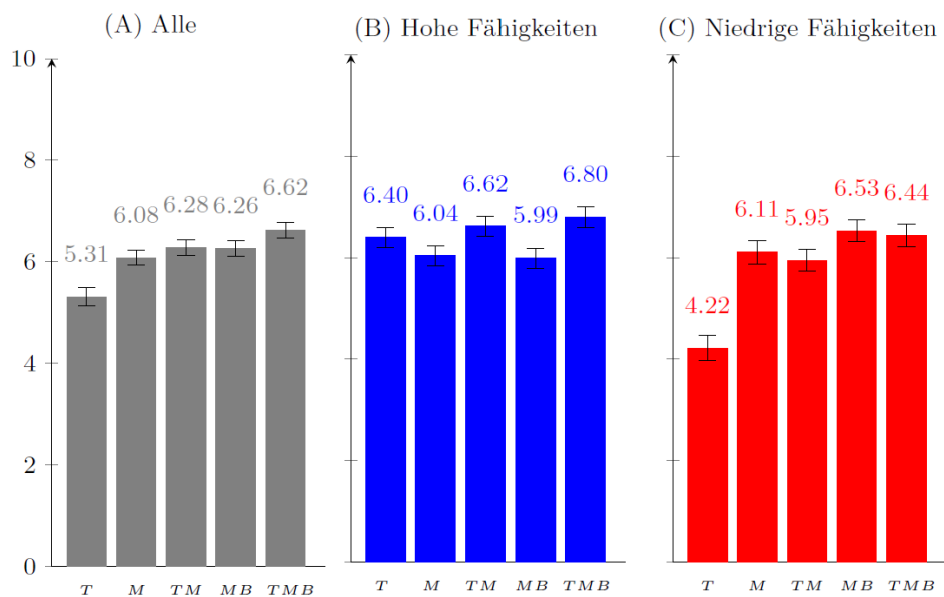
In Bezug auf Hypothese 1 folgern wir also, dass:

**Resultat 1:** Die Anstrengung von Ärztinnen und Ärzten hängt von ihren Fähigkeiten und dem Design des Rankings ab

- a) Die Anstrengung erhöht sich mit der Anzahl der erreichbaren Ranggrenzen
- b) Ranggrenzen, die nicht erreicht werden können, neigen dazu, die Anstrengung zu senken
- c) Ranggrenzen, die nicht unterschritten werden können, beeinflussen die Anstrengung nicht



**Abbildung 3: Anstrengung unter verschiedenen Rankings**



Mittlere Anstrengung für (A) alle Ärztinnen und Ärzte, (B) Ärztinnen und Ärzte mit hohen Fähigkeiten, (C) Ärztinnen und Ärzte mit niedrigen Fähigkeiten

Betrachtet man die beiden Gruppen zusammen, siehe Panel A von Abbildung 3, so findet man, dass die Anstrengung mit der Anzahl der Ranggrenzen pro Ranking steigt. Das heißt, die Rankings mit zwei Ranggrenzen (TM-Ranking und MB-Ranking) führen zu signifikant höherer Anstrengung als die Rankings mit einer Ranggrenze (T-Ranking und M-Ranking). Zudem führt das granulare TMB-Ranking mit allen Ranggrenzen zu signifikant mehr Anstrengung als alle Rankings mit zwei Ranggrenzen. Hieraus folgt, dass wenn für jede Gruppe eine erreichbare Ranggrenze im Ranking ist, der motivierende Effekt von zusätzlichen Ranggrenzen immer den demotivierenden Effekt überwiegt. Das in der Praxis populäre Ranking, welches nur das Spitzenenergebnis hervorhebt, ist hingegen sehr demotivierend für Ärztinnen und Ärzte mit niedrigen Fähigkeiten und führt zu der geringsten Anstrengung.

In Bezug auf Hypothese 2 folgern wir demnach, dass:

**Resultat 2:** Die mittlere Ranggrenze, die alle erreichen können, motiviert alle unabhängig von der Fähigkeit. Insgesamt motivieren zusätzliche Ranggrenzen, solange für jeden Fähigkeitstyp mindestens schon eine erreichbare Ranggrenze im Ranking enthalten ist.

### Sollten Klinikleitungen Rankings überhaupt einführen?

Der Vergleich zwischen den verschiedenen Rankings hat gezeigt, dass das granulare Ranking zu der höchsten Anstrengung führt. Dennoch bleibt die Frage, ob Krankenhäuser sich mit der Einführung von Rankings besser oder schlechter stellen. Tabelle 1 fasst die Behandlungsergebnisse zusammen.

Tabelle 1 bestätigt die widersprüchlichen Befunde der bisherigen Literatur. Insgesamt kann die Einführung eines Rankings zu mehr und weniger Anstrengung führen. Das TMB-Ranking, das alle erreichbaren Ranggrenzen kombiniert, steigerte die Anstrengung um 5 % im Vergleich zur Baseline ohne Rankings. Hingegen reduzierte das Ranking, das nur das Spitzenenergebnis belohnt (T-Ranking), die Anstrengung drastisch um 16 %. Andere Rankings hatten keinen signifikanten Einfluss auf die Anstrengung im Vergleich zur Baseline ohne Ranking über die Fähigkeitsstufen hinweg.

**Tabelle 1: Anstrengungsänderung unter Rankings relativ zur Baseline ohne Ranking in %**

	<i>Alle Fähigkeiten</i>	<i>Hohe Fähigkeiten</i>	<i>Niedrige Fähigkeiten</i>
<i>T-Ranking</i>	–15,43***	7,75***	–36,22***
<i>M-Ranking</i>	–3,21	1,72	–7,64*
<i>TM-Ranking</i>	0,09	11,48***	–10,13***
<i>MB-Ranking</i>	0,23	0,96	–1,29
<i>TMB-Ranking</i>	5,43***	14,55***	–2,75

Notiz: Die Tabelle gibt die prozentuale Änderung der Anstrengung wieder. \*, \*\*, \*\*\* implizieren eine signifikante Änderung im Vergleich zur Baseline ohne Ranking bei einem Signifikanzniveau von 10%, 5% und 1%.

Betrachtet man die verschiedenen Fähigkeiten, so findet man, dass Ärztinnen und Ärzte mit hohen Fähigkeiten von keinem der fünf Rankings demotiviert sind. In der Tat ist ihre Anstrengung unter jedem Ranking im Mittel höher als ohne Feedback – für Rankings die die obere Ranggrenze (T) enthalten ist dies signifikant. Auf der anderen Seite sind Ärztinnen und Ärzte mit niedrigen Fähigkeiten nie von Rankings motiviert. In der Tat ist ihre Anstrengung unter jedem Ranking im Mittel niedriger – für Rankings in denen nicht sowohl die mittlere (M) wie auch die untere Ranggrenze (B) enthalten ist, ist dies signifikant.

## Fazit

In unserer Studie haben wir untersucht, wie Rankings die Anstrengung von Ärztinnen und Ärzten mit unterschiedlichen Fähigkeiten beeinflussen. Ziel war es, herauszufinden, wie Rankings gestaltet werden sollten, um die Motivation am besten zu fördern, ohne einzelne zu demotivieren.

Unsere Studie findet eine motivierende Wirkung erreichbarer individueller Ziele in Form von Ranggrenzen. Rankings mit mehreren Ranggrenzen, die verschiedene Fähigkeitstypen ansprechen, erzielen die größten Effekte. Insbesondere das TMB-Ranking, das Ranggrenzen für hohe, mittlere und niedrige Behandlungsergebnisse kombiniert, steigerte die Anstrengung der Ärztinnen und Ärzte signifikant zur Situation ohne Rankings und im Vergleich zu anderen Rankings. Ärztinnen und Ärzte mit hohen Fähigkeiten wurden durch Spitzenplätze motiviert, während Ärztinnen und Ärzte mit niedrigen Fähigkeiten durch erreichbare Ziele nicht demotiviert wurden.

Wir finden außerdem ein hohes Risiko von Demotivation durch unerreichbare Ziele. Rankings, die sich nur auf Spitzenplätze konzentrieren (T-Ranking), führten zu einem deutlichen Rückgang der Anstrengung bei Ärztinnen und Ärzten mit niedrigen Fähigkeiten (bis zu 36 % weniger Anstrengung). Dies zeigt, dass Rankings, die aus Zielen bestehen, welche für die meisten unerreichbar sind, kontraproduktiv sein können.

Insgesamt ist die individuelle Anpassung entscheidend. Die Gestaltung eines Rankings sollte an die Fähigkeiten eines Teams angepasst werden. Beispielsweise kann in einem Team mit wenigen Ärztinnen und Ärzten mit hohen Fähigkeiten eine weniger exklusive höchste Ranggrenze sinnvoll sein, um breitere Motivation zu fördern.

Unsere Behandlungsergebnisse können direkt auf die Gestaltung von Leistungsfeedback in klinischen Teams angewandt werden. Zum Beispiel könnten Klinikleitungen regelmäßige Bewertungen einführen, die Rankings mit erreichbaren Ranggrenzen enthalten. Dies könnte die Qualität der Versorgung steigern, indem es sowohl Spitzenleistungen fördert als auch eine breitere Motivation sicherstellt.

Als letztes ist hervorzuheben, dass unser Experiment vereinfachte Szenarien simuliert. Reale Arbeitsbedingungen und psychologische Faktoren, wie Teamdynamiken, wurden nicht vollständig erfasst. Aufgrund von Rückmeldungen von Klinikleitungen, die das Setting als praxisrelevant einschätzen, glauben wir, dass unser Experiment die wesentlichen Mechanismen identifizieren kann.

Im Allgemeinen dienen Experimente als exzellenter Windkanal, um Managementinterventionen im Gesundheitswesen zu testen, da sie es ermöglichen, menschliches Entscheidungsverhalten unter kontrollierten Bedingungen (realitätsnah) zu beobachten und systematische Abweichungen von theoretischen Annahmen zu identifizieren. Insbesondere bieten sie die Flexibilität, verschiedene Einflussfaktoren gezielt zu testen, um kausale Zusammenhänge zwischen Verhalten und Anreizen zu analysieren (siehe auch „Einsatz verhaltensökonomischer Interventionen zur Verbesserung ärztlicher Entscheidungen“ im Deutschen Ärzteblatt)<sup>9</sup>.

## Quellenverzeichnis

- 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, Iriberry 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.
- Charness G, Masclet D, Villeval MC (2014) The dark side of competition for status. *Management Science* 60(1):38–55.
- Edelman B, Larkin I (2015) Social comparisons and deception across workplace hierarchies: Field and experimental evidence. *Organization Science* 26(1):78–98.
- 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, Kissova 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.
- Hallek M, Ockenfels A, Wiesen D (2022) Einsatz verhaltensökonomischer Interventionen zur Verbesserung ärztlicher Entscheidungen. *Deutsches Ärzteblatt* 119(38):633–639.
- 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.
- Huesmann K, Ndiaye YS, Waibel C, Wiesen D (forthcoming) Peer Feedback in Healthcare: How Ranking Designs and Ability Affect Physician Effort. *Management Science*.
- 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. *Quarterly Journal of Economics* 129(1):105–149.
- 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.
- Niewoehner RJ, Staats BR (2022) Focusing provider attention: An empirical examination of incentives and feedback in flu vaccinations. *Management Science* 68(5):3680–3702.
- Singh M, Zureich J (2023) Do physicians improve more from positive or negative feedback? Mimeo.

---

<sup>9</sup> Hallek et al. (2022)

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

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.

---

#### IMPRESSUM

Abbildung S.1:  
<https://pixabay.com/de/photos/behandlung-krankenhaus-klinik-4099432/>

#### Autorenkontakt:

Yero S. Ndiaye  
Lehrstuhl Axel Ockenfels an der  
Universität zu Köln  
Universitätsstraße 22a, Gebäude  
102, Raum 4.203  
50923 Köln  
[ndiaye@wiso.uni-koeln.de](mailto:ndiaye@wiso.uni-koeln.de)

#### Herausgeber:

Institut für Wirtschaftspolitik  
an der Universität zu Köln  
Pohligstr. 1  
50969 Köln  
Tel. 0221 / 470-5347  
Fax 0221 / 470-5350  
[iwp@wiso.uni-koeln.de](mailto:iwp@wiso.uni-koeln.de)

#### Redaktion und V.i.S.d.P.:

Prof. Dr. Steffen J. Roth  
Tel. 0221 / 470-5348  
[steffen.roth@wiso.uni-koeln.de](mailto:steffen.roth@wiso.uni-koeln.de)