Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Do Physicians Improve More from Positive or Negative Feedback?

Manasvini Singh, Jacob Zureich

To cite this article:
Manasvini Singh, Jacob Zureich (2024) Do Physicians Improve More from Positive or Negative Feedback?. Management Science

Published online in Articles in Advance 02 Sep 2024

. https://doi.org/10.1287/mnsc.2023.01340

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

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–25
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Do Physicians Improve More from Positive or Negative Feedback?

Manasvini Singh,[a] Jacob Zureich[b,*]

[a] Social and Decision Sciences Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [b] Accounting Department, College of Business, Lehigh University, Bethlehem, Pennsylvania 18015

*Corresponding author

Contact: manasvinisin@umass.edu, https://orcid.org/0000-0002-0710-1117 (MS); jaz423@lehigh.edu, https://orcid.org/0000-0003-1529-0920 (JZ)

Received: May 2, 2023
Revised: January 26, 2024; February 23, 2024
Accepted: February 29, 2024
Published Online in Articles in Advance: September 2, 2024

https://doi.org/10.1287/mnsc.2023.01340

Abstract. We use clinical data on more than 240,000 surgeries and quasi-experimental methods to examine how physicians respond to the surprise release of a performance "report card." Such feedback interventions are commonly used to encourage physicians to improve performance yet show limited evidence of success. Our results show that these limited effects mask heterogeneous behavioral responses to feedback valence. In particular, physicians improve more from positive feedback than from negative feedback, with negative feedback even reducing performance for a nontrivial share of patients. Experiments with laypersons replicate these results and show that struggles with negative feedback can be mitigated by giving incentives directly tied to improvement and by adding qualitative information that helps individuals interpret past performance. These results are consistent with behavioral models that suggest cognitive and emotional difficulties limit how well individuals use negative feedback. Thus, feedback interventions in healthcare should be carefully designed to mitigate these counterproductive behavioral responses.

History: Accepted by Ranjani Krishnan, accounting.
Funding: We acknowledge generous funding provided by Tilburg University and the University of Massachusetts Amherst to run the laboratory experiments.
Supplemental Material: The online appendices and data files are available at https://doi.org/10.1287/mnsc.2023.01340.

Keywords: accounting • healthcare • healthcare: treatment • learning • feedback

1. Introduction

The past decade in healthcare has seen a dramatic rise in the use of outcome controls to hold physicians accountable for patient care (Hackbarth et al. 2008, VanLare et al. 2012, MacLean et al. 2018). Performance feedback on patient outcomes is a key prong in this push toward accountability, with the Affordable Care Act declaring "Measurement and Feedback" to be an essential part of their strategy for quality improvement (AHRQ 2017). In response, a growing number of nonprofit and governmental organizations have begun releasing physician report cards, which publicly disclose ratings of individual physicians' performance. Although these report cards have been met with mixed reviews and vigorous debate, emerging literature suggests that they lead to moderate improvements in health outcomes on average (Hofer et al. 1999, Fung et al. 2008, Prang et al. 2021). In this study, we show that these average improvements are more nuanced than previously thought. Using a large clinical data set and a series of experiments, our findings suggest that report cards can have widely varying effects depending on whether they give physicians positive or negative performance feedback. The report cards even lead to systematically worse patient outcomes, that is, higher rates of patient mortality and hospital readmission, in a nontrivial portion of situations.

Understanding how individuals respond to new performance information has been a longstanding question in accounting (Frederickson 1992, Kaplan et al. 1996, Hannan et al. 2008, Casas-Arce et al. 2017, Anderson and Kimball 2019, Manthei et al. 2022) and in the social sciences more broadly (Kluger and DeNisi 1996, Levitt et al. 2013, Song et al. 2018). There has been particular interest on the distinction between negative and positive feedback (Eil and Rao 2011, Lefebvre et al. 2017, Loftus and Tanlu 2018, Eskreis-Winkler and Fishbach 2019, Zimmermann 2020, Erickson et al. 2022). We define positive and negative performance feedback as new information indicating that past performance was better or worse than previously thought. Notably, feedback valence (positive versus negative) is distinct

from ability, and both bottom and top performers can learn they were better or worse than they previously thought.

Conventional wisdom and some academic theories suggest that failure teaches more than success (Locke and Latham 1990, Baumeister et al. 2001, Cannon and Edmondson 2005, Syed 2015). However, more recent theory and empirical evidence challenge this thinking. Although negative feedback may potentially be more useful, people often struggle to use it effectively because they find it emotionally and cognitively challenging to accept (Nease et al. 1999, Lefebvre et al. 2017, Eskreis-Winkler and Fishbach 2019, Chambon et al. 2020, Erickson et al. 2022). Negative feedback can also be dejecting whereas positive feedback fosters intrinsic motivation and confidence (Deci et al. 1999, Azmat and Iriberri 2010). Building on this prior research, we examine whether these effects also explain the way physicians respond to new performance information: Are physicians motivated and capable enough to use negative feedback effectively, or do they also exhibit adverse behavioral responses and improve more from positive feedback?

We assess this question using the surprise release of a nationwide surgeon report card in 2015. The report card was released online by ProPublica (a highly regarded nonprofit investigative journalism outlet) and provided new outcome feedback by evaluating all surgeons nationwide on their patient outcomes adjusted for uncontrollable factors like patient risk and hospital quality. We link this report card data to clinical data on all inpatient encounters in the state of Florida over five years and examine how objective indicators of physician performance (rates of 30-day readmission and in-hospital mortality) change after the release of the report card. Physicians have no obligation to respond to this report card from an independent nonprofit like ProPublica. Yet it may still induce a physician response because it provides (i) new individual performance information by adjusting for uncontrollable factors (i.e., reducing noise), and (ii) new relative performance information by ranking surgeons compared with peers.

Our first analysis isolates the causal effects of feedback valence from other factors (e.g., physician ability) by using a regression-discontinuity design. The ProPublica report card placed surgeons into one of three color-coded zones based on their performance, with the top, middle, and bottom performers placed in the green, yellow, and red zones, respectively. Our analysis leverages the arbitrary cutoffs used for these zones as discontinuities. The results show that surgeons who barely make the green zone improve anywhere from 4.5 to 9.5 percentage points more than surgeons in the yellow zone who barely miss making the green zone. However, there are no such effects with the yellow and red zones. Thus, receiving positive versus less positive feedback (green versus yellow) improves performance but receiving negative versus less negative feedback (red versus yellow) has negligible effects.

We next use a difference-in-difference (DiD) design using the two measures of feedback valence - one based on individual and the other on relative performance feedback. We first show support for the parallel trends assumption: physicians who ultimately receive negative versus positive feedback were trending similarly before the report card release. However, after the release, those who receive positive feedback exhibit a strong upward trend in performance compared with those who receive negative feedback. Specifically, physicians who receive positive individual performance feedback subsequently reduce their patient complication rates by 2.5% more than those who receive negative feedback. With the relative measure, this difference-in-difference estimate of valence is 1.1%. These effects hold regardless of physicians' performance in the preperiod and, interestingly, negative feedback leads to a reduction in performance for previously high-performing physicians.[1] These findings are not consistent with mean reversion and additional analyses suggest that physicians who receive positive feedback do not improve more because they see healthier patients or have stronger financial incentives. There is also no evidence that demand effects by patients are driving the results.

To more definitively assess whether the effects of valence are due to beneficial effects of positive feedback, detrimental effects of negative feedback, or both, we also conduct an additional within-physician analysis using the nontreated surgeries (i.e., surgeries not graded by the report card) as controls. Compared with control surgeries (which, as expected, were not affected by the report card), positive feedback on treated surgeries leads to increased performance while negative feedback leads to decreased performance (though statistically significant only with individual performance feedback). Taken together, there is strong evidence that positive feedback is beneficial for surgeon performance and moderate evidence that negative feedback is detrimental to their performance.

Overall, it is concerning that physicians fail to use negative feedback effectively in a setting with severe consequences on patient health. Even more concerning are the systematic reductions in patient outcomes that can result from negative feedback. Although we cannot precisely pinpoint mechanisms, the results support behavioral theories in which people protect their ego by ignoring and discounting the validity of negative feedback. For example, surgeons who receive low report card scores might rationalize them by discounting the methods used to construct the scores (Nease et al. 1999, Gnepp et al. 2020). The observed reductions in patient outcomes in some cases are consistent with behavioral theories about the cognitive and motivational challenges of using negative feedback effectively. For example, physicians who do

accept negative feedback may become demotivated (Deci et al. 1999) or channel their efforts into less productive task strategies (Hannan et al. 2008).

To learn more about these effects, we also ran a series of online experiments with participants from Prolific and Amazon's Mechanical Turk. The experiments provide clean identification by manipulating feedback valence and including control conditions in which no feedback is shown to participants. The results are strikingly consistent with those of the surgeon analysis. The only substantive difference is that the experiments provide somewhat stronger evidence that negative feedback is detrimental and weaker evidence that positive feedback is beneficial (discussed more in Section 4.1.2). However, in both analyses, the directional effects are consistent: positive feedback helps and negative feedback harms.

Finding consistent results with both surgeons and laypersons points to generalized underlying behavioral factors. Results from our first experiment further show that the differential response to negative and positive relative performance feedback are primarily driven by more competitive individuals. This effect of competitiveness may shed light on the physician results because medicine is a particularly competitive field and competitiveness can be an important driver of physician behavior (Liao et al. 2016, Meeker et al. 2016, Murphy 2018, Moon 2021). Results from the second experiment further show that the differential response to positive and negative feedback can be reduced by (i) giving incentives explicitly tied to improvement (in addition to incentives for overall performance) or (ii) giving qualitative information on how to interpret feedback. Thus, the poor response to negative feedback increases with competitiveness and may stem from both the added motivational and cognitive challenges associated with ego-damaging performance feedback.

Our study contributes to research on performance feedback and the growing debate on the use of physician report cards in healthcare (Andrabi et al. 2017, Eyring 2020, Gallani et al. 2020). We extend prior research on the potential lackluster effects of performance feedback, especially negative feedback, to a setting with physicians making consequential decisions. We also respond to recent calls in the social sciences for a "heterogeneity revolution" (Bryan et al. 2021). Moving beyond a focus on average effects provides a more complete picture of how physician report cards affect behavior and shows that the mild on average effects observed in previous research may have been understating the true extent of physician response. Finding that physicians fail to use negative feedback effectively should be concerning to patients, hospitals, and regulators. The finding that report cards increase the performance of some physicians but decrease it for others also suggests that these disclosures increase interphysician variation in care that is unrelated to patient characteristics: an important concern in healthcare (Fisher et al. 2003, Wennberg 2004).

From a practical perspective, our focus on heterogeneities sheds light on how to improve feedback interventions in healthcare. Our results suggest that efforts to improve report cards and other forms of feedback should center on how to help physicians use negative feedback more effectively (Edmondson 2011). Results from our experiment indicate that pairing outcome feedback with improvement incentives or qualitative information about how to use the feedback are both promising avenues for practitioners to consider. Prior research suggests a number of other avenues such as encouraging a promotion rather than prevention focus and focusing feedback on future actions rather than past performance (Van-Dijk and Kluger 2004, Gnepp et al. 2020). Report card providers should also consider bolstering the perceived legitimacy of the report cards. Physicians strongly criticize the underlying methods used to create report card ratings, and people are more likely to discount negative feedback when they judge its source to be inaccurate or invalid (Levy and Williams 2004, Loftus and Tanlu 2018, Gnepp et al. 2020).

The paper proceeds as followers. Section 2 develops theory. Section 3 presents results of the physician analysis, whereas Section 4 presents results of the experiments. Section 5 concludes.

2. Theoretical Development

The United States lags behind its Organisation for Economic Co-operation and Development contemporaries in healthcare quality, having higher rates of infant mortality and lower life expectancy, despite having one of the most expensive healthcare systems in the world (Anderson and Hussey 2001, Tikkanen and Abrams 2020). In an attempt to redress these deficiencies, the Affordable Care Act implemented a slew of changes with the goal of increasing quality and reducing costs. The first of the nine levers it identified as central to achieving these goals is "Measurement and Feedback," which encourages stakeholders (such as healthcare organizations) to "provide performance feedback to plans and providers to improve care" (AHRQ 2017). As a result, physician report cards, which publicly disclose performance information on specific physicians or hospitals, have witnessed a sharp rise in popularity in the last decade (Christianson et al. 2010, Shi et al. 2017).

Performance disclosures have been successful in motivating improvements across a wide range of domains such as in social responsibility and restaurant hygiene (Jin and Leslie 2003, 2009; Friesen et al. 2012; Evans 2016; Leuz and Wysocki 2016; Christensen et al. 2017; Chen et al. 2018). However, in healthcare, reports cards have seen mixed success. Although a primary

goal of report cards was to incentivize improvement through reputational effects, these incentives are often negligible because report card scores "have had minimal impact on consumer choices of providers" (Uhrig and Short 2002, Hirth et al. 2003, Epstein 2010, Grabowski and Town 2011, Jung et al. 2011, Sinaiko et al. 2012, Scanlon et al. 2015).[2] The concluding sentence of a recent New York Times article nicely summarizes these issues (Jauhar 2015, p. 27): "It would appear that doctors, not patients, are the ones focused on doctors' grades, and their focus is distorted and blurry at best." Thus, although physicians have financial incentives to perform well and have been shown to respond to those incentives (Jacobson et al. 2010, Clemens and Gottlieb 2014, Khullar et al. 2015), there is little evidence that the release of physician report cards amplifies these incentives in beneficial ways (Kolstad 2013).[3]

Despite these issues, accumulating evidence suggests that report cards do lead to mild performance improvements on average (Fung et al. 2008, Dunt et al. 2018, Prang et al. 2021). Even when report cards do not increase the extrinsic incentive to improve, they can still be effective because physicians intrinsically care about improving (Kolstad 2013). Although these effects are promising, prior studies often focus on organizational-level rather than individual-level report cards (Fung et al. 2008). Moreover, research on performance feedback in general paints a more complicated picture, showing that the provision of additional performance information can be both beneficial, neutral, or detrimental depending on features of the task, feedback, and individual (Balcazar et al. 1985, Kluger and DeNisi 1996, Hannan et al. 2008, Casas-Arce et al. 2017, Loftus and Tanlu 2018, Anderson and Kimball 2019, Eyring et al. 2021, Erickson et al. 2022). Thus, many open questions remain such as when and for whom the positive effects of giving physicians performance feedback are most likely to occur. To gain a more complete understanding of whether report cards are achieving their objectives and how they can be made more effective in the future, we move beyond a focus on average effects to examine key heterogeneities in individual physician response.

We focus specifically on whether physicians respond differently when they receive positive and negative performance feedback from the report card. Positive feedback is new information indicating that performance was better than previously thought whereas negative feedback is new information indicating that performance was worse than previously thought. Report cards can be an important source of positive or negative outcome feedback because they give physicians new relative performance information, and they filter out noise by adjusting prior performance for uncontrollable factors. Importantly, both higher- and lower-quality physicians can receive positive or negative feedback. For example, consider high-quality physicians who expect to be ranked better than 90% of their peers. These physicians would receive negative feedback if they learn that they are only ranked better than 75% of their peers. As will be described, we use various techniques to isolate the effect of feedback valence from the effects of physician characteristics like quality.

The topic of how individuals respond to positive and negative feedback has received considerable attention in the popular press and across a wide range of fields (Kluger and DeNisi 1996, Bailey 2014, Lefebvre et al. 2017, Loftus and Tanlu 2018, Eskreis-Winkler and Fishbach 2019, Chambon et al. 2020, Erickson et al. 2022). Many studies have asked a simple but elusive question: does positive or negative feedback work better? Theory suggests an important role for both types of feedback (Hogarth et al. 1991). For example, both negative and positive feedback can help individuals refine their task strategies by providing valuable information about the efficacy of past actions (Hannan et al. 2008). Additionally, social comparison theory suggests that relative performance information will increase effort, regardless of valence (Festinger 1954, Tafkov 2012, Kolstad 2013).

Some theories suggest that negative feedback should induce greater improvements than positive feedback because "bad is stronger than good" (Baumeister et al. 2001). For example, individuals spend more time processing and thinking about negative information than positive information and loss aversion is predicated on the idea that people care more about negative than positive events (Klinger et al. 1980, Gilovich 1983, Weiner 1985, Kahneman and Tversky 2013). Other arguments suggests that negative feedback results in greater behavioral change because it indicates the need for adjustment, whereas positive feedback merely reinforces the continuation of previous behavior (Carver and Scheier 2001, Lourenço 2016).

However, there is an emerging consensus from empirical research that positive feedback is more likely to increase motivation and to be used effectively for learning (Eil and Rao 2011, Kc et al. 2013, Lefebvre et al. 2017, Eskreis-Winkler and Fishbach 2019, Chambon et al. 2020). Negative feedback often has less impact because people are motivated to reject it to maintain a positive self-image (Eil and Rao 2011, Loewenstein and Molnar 2018, Elder et al. 2022). Because of these motivated beliefs, compared with people who receive positive feedback, those who receive negative feedback are more likely to (i) be critical of its accuracy (Ruzzene and Noller 1986, Gnepp et al. 2020), (ii) attribute it to external forces like bad luck (Grossman and Owens 2012, Dorfman et al. 2019, Chambon et al. 2020), and forget about it over time (Zimmermann 2020). Even when individuals do accept negative feedback as valid, it can be more difficult to act on than positive feedback. Although positive feedback indicates what to do (double-down on previous strategies),

negative feedback only indicates what not to do, leaving the best path forward ambiguous and causing individuals to adopt ineffective strategies (Hogarth et al. 1991, Hannan et al. 2008). Furthermore, whereas accepting the validity of negative feedback can lead to anger, stress, lack of confidence, and dejection, positive feedback can be energizing and increase intrinsic motivation (Harackiewicz 1979, Deci et al. 1999, Nease et al. 1999, Audia and Locke 2003, Compte and Postlewaite 2004, Hattie and Timperley 2007, Swift and Peterson 2018, Fong et al. 2019, Erickson et al. 2022, Awaysheh et al. 2023).[4] Overall, evidence suggests that positive feedback generally improves performance and negative feedback often has little, or even negative, effects on performance. Thus, we make the following directional prediction.

Hypothesis 1. *Physician report cards lead to greater performance improvements for physicians who receive positive feedback than for those who receive negative feedback.*

This hypothesis is not without tension. Prior research suggests that the adverse effects associated with negative feedback are less likely to occur for experts and individuals with high self-esteem: both of which plausibly characterize physicians (Brockner et al. 1987, Ilies et al. 2007, Brown 2010, Finkelstein and Fishbach 2012). Furthermore, negative feedback has been shown to be more effective when individuals are highly committed to performing their task well (Fishbach et al. 2010). Thus, physicians, who likely take pride in their work and whose actions have severe consequences for patient health, may be motivated and capable enough to use negative feedback effectively. For these reasons, physicians may even respond more to negative feedback than to positive feedback and it is important to provide causal evidence within this domain.

3. Physician Analysis

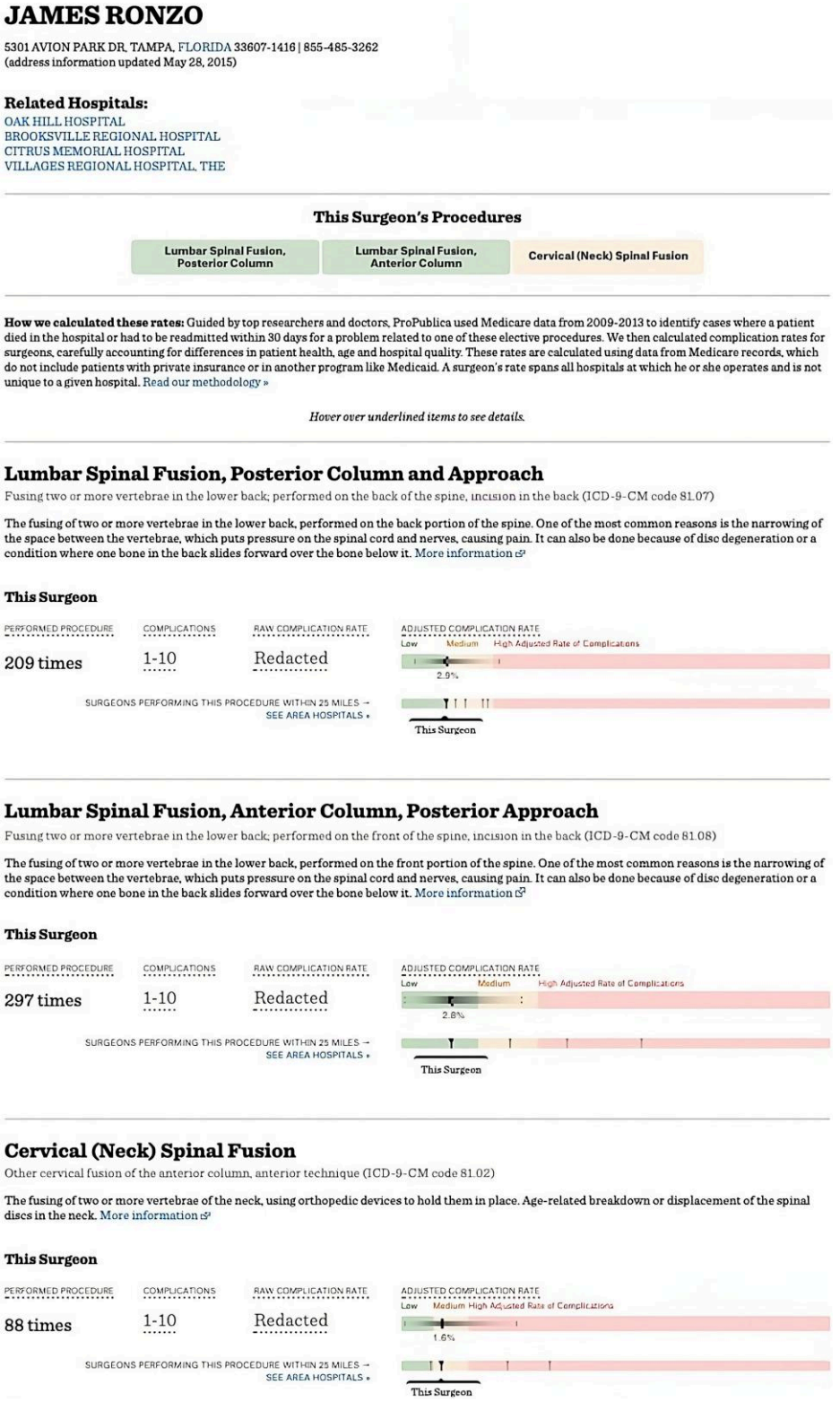
Setting: Physician Report Card. Our study investigates the response of physicians to the unexpected release of an online "Report Card" by ProPublica (a prominent pro-consumer nonprofit) in July 2015, which evaluated surgeons based on patient outcomes from eight low-risk elective surgeries performed on all Medicare patients in the United States between 2009 and 2013.[5] These eight surgeries are hereafter referred to as "key procedures." Physicians, evaluated on (i) in-hospital mortality and (ii) 30-day readmission, were informed of their "adjusted complication rate" (hereafter, report card complication rate) for each procedure, which adjusted their performance for hospital and patient factors, as well as luck at extremes. Surgeons' rates were categorized into green, yellow, or red zones, indicating above average, average, and below average complication rates, respectively, for a given procedure. The report card also showed peer comparisons with surgeons performing the same procedures within a 25-mile radius of their primary business address. Figure 1 shows an example of a surgeon's report card.

This report card's release has several key aspects for consideration. First is whether this release constitutes an "exogenous" shock to physician information. ProPublica, known for impactful journalism, typically does not preannounce projects to create more impact on release, thus differing on this dimension from healthcare entities encouraged by the ACA to provide physician feedback. This approach is evident in the case of the surgeon report card; a search in the Wayback Machine, which archives more than 800 billion web pages, reveals no saved copies before August 1, 2015 (Figure A.1 in the Online Appendix). This indicates the absence of the web page before this date, implying it was not available earlier. Consequently, it seems improbable that surgeons were aware of the report card prior to its release, leading us to consider the release as an unexpected event that provided physicians with new performance information.

The second point to consider is whether the report card's single release in 2015 can be expected to have a measurable impact on physician performance, despite its lack of annual updates. We argue that it can. The report card's release sparked significant discussion in the medical community about the ethics and implications of physician scoring systems, drawing attention from physicians (Engelberg and Pierce 2015, Rosenbaum 2015, Friedberg et al. 2016, Jaffe et al. 2016). It was notable for providing a permanent online record of a one-time evaluation of physician performance. Although ProPublica did not communicate plans for future updates, physicians might have reacted in the hope of bettering their scores in potential future editions.[6]

The final aspect to discuss is whether the report card offered new information to physicians. One might think that physicians are already aware of their performance through internal hospital feedback. However, this is unlikely in inpatient settings due to the complex relationship between hospitals and many physicians, particularly those in our study who perform elective surgeries (Goldsmith et al. 2016, Scott et al. 2017). Many physicians are not directly employed by hospitals, so they do not receive typical employer-employee feedback. Additionally, hospitals cannot access a physician's performance data from other hospitals due to HIPAA restrictions, limiting their ability to provide comprehensive feedback. The ProPublica Report Card, using nationwide Medicare surgery data, offers a more accurate and broader performance evaluation than any single hospital could. Our study also uses a hospital fixed effect in the analysis, ensuring within-hospital comparisons and negating the need to consider variation in internal feedback across hospitals. Moreover, if hospitals already provide similar information, then

Figure 1. Report Card Example



JAMES RONZO

5301 AVION PARK DR, TAMPA, FLORIDA 33607-1416 | 855-485-3262
(address information updated May 28, 2015)

Related Hospitals:
OAK HILL HOSPITAL
BROOKSVILLE REGIONAL HOSPITAL
CITRUS MEMORIAL HOSPITAL
VILLAGES REGIONAL HOSPITAL, THE

This Surgeon's Procedures

Lumbar Spinal Fusion, Posterior Column | Lumbar Spinal Fusion, Anterior Column | Cervical (Neck) Spinal Fusion

How we calculated these rates: Guided by top researchers and doctors, ProPublica used Medicare data from 2009-2013 to identify cases where a patient died in the hospital or had to be readmitted within 30 days for a problem related to one of these elective procedures. We then calculated complication rates for surgeons, carefully accounting for differences in patient health, age and hospital quality. These rates are calculated using data from Medicare records, which do not include patients with private insurance or in another program like Medicaid. A surgeon's rate spans all hospitals at which he or she operates and is not unique to a given hospital. Read our methodology »

Hover over underlined items to see details.

Lumbar Spinal Fusion, Posterior Column and Approach

Fusing two or more vertebrae in the lower back, performed on the back of the spine, incision in the back (ICD-9-CM code 81.07)

The fusing of two or more vertebrae in the lower back, performed on the back portion of the spine. One of the most common reasons is the narrowing of the space between the vertebrae, which puts pressure on the spinal cord and nerves, causing pain. It can also be done because of disc degeneration or a condition where one bone in the back slides forward over the bone below it. More information

This Surgeon

PERFORMED PROCEDURE	COMPLICATIONS	RAW COMPLICATION RATE	ADJUSTED COMPLICATION RATE
209 times	1-10	Redacted	Low / Medium / High Adjusted Rate of Complications; 2.9%

SURGEONS PERFORMING THIS PROCEDURE WITHIN 25 MILES →
SEE AREA HOSPITALS »
This Surgeon

Lumbar Spinal Fusion, Anterior Column, Posterior Approach

Fusing two or more vertebrae in the lower back, performed on the front of the spine, incision in the back (ICD-9-CM code 81.08)

The fusing of two or more vertebrae in the lower back, performed on the front portion of the spine. One of the most common reasons is the narrowing of the space between the vertebrae, which puts pressure on the spinal cord and nerves, causing pain. It can also be done because of disc degeneration or a condition where one bone in the back slides forward over the bone below it. More information

This Surgeon

PERFORMED PROCEDURE	COMPLICATIONS	RAW COMPLICATION RATE	ADJUSTED COMPLICATION RATE
297 times	1-10	Redacted	Low / Medium / High Adjusted Rate of Complications; 2.8%

SURGEONS PERFORMING THIS PROCEDURE WITHIN 25 MILES →
SEE AREA HOSPITALS »
This Surgeon

Cervical (Neck) Spinal Fusion

Other cervical fusion of the anterior column, anterior technique (ICD-9-CM code 81.02)

The fusing of two or more vertebrae of the neck, using orthopedic devices to hold them in place. Age-related breakdown or displacement of the spinal discs in the neck. More information

This Surgeon

PERFORMED PROCEDURE	COMPLICATIONS	RAW COMPLICATION RATE	ADJUSTED COMPLICATION RATE
88 times	1-10	Redacted	Low / Medium High Adjusted Rate of Complications; 1.6%

SURGEONS PERFORMING THIS PROCEDURE WITHIN 25 MILES →
SEE AREA HOSPITALS »
This Surgeon

Notes. This figure shows a real example of the Propublica report card for an example surgeon. This surgeon performed one of the key surgeries (lumber spinal fusion) 209 times with an adjusted complication rate of 2.9%, which scored in the green zone.

the report card should not change physician behavior, making it less likely that we observe any effects.

Clinical Data. We link data on physicians from the ProPublica report card to two sources of inpatient clinical data from the state of Florida: (i) Florida's State Inpatient Utilization Database from the Healthcare Cost and Utilization Project (HCUP) from 2012 to 2016, and (ii) the Agency For Healthcare Administration (AHCA) Inpatient Database from the state of Florida from 2012 to 2016.[7] These fully deidentified datasets (i.e., data sets that have no identifying information on the patient, such as date of birth, date of patient hospital admission, etc.) are collected by government for quality and research purposes and provide data on 100% of all inpatient encounters that occurred in all hospitals in the state of Florida. Both data sets contain the exact same raw data from hospitals with differences only due to postdata collection processing and reporting standards (because the HCUP is reported by the federal government, and the AHCA is reported by the state). However, we need both data sets because only the HCUP reports 30-day readmission and only the AHCA reports the physician of record (via National Provider Identifiers) for each encounter, both of which are critical to our analysis. Despite de-identification, they contain a host of variables on patient characteristics, such as socio-demographics, quarter of admission/discharge, and clinical (diagnosis and procedure) codes.

For each data set, we limit the sample to encounters with patients who had an elective procedure from 2012 to 2016 and who were (i) over 50, (ii) not transferred in from another hospital, and (iii) not admitted to the hospital through the emergency department (ED). We merge the two data sets on the following variables: year, quarter, age, sex, race, length of stay, whether the encounter was on a weekday, principal diagnosis, principal procedure, Elixhauser comorbidities, zip code of residence, hour of admission, and hour of discharge. This led to an 80% perfect match for all records in HCUP (for nonmatches, one of the aforementioned variables was missing in either the HCUP or AHCA data), resulting in a sample of 1.16 million patient encounters.

We then restrict this sample to encounters with a key procedure for their principal procedure by physicians who were provided report card feedback from Propublica, and then finally, to those physicians who performed a key procedure every quarter from 2012 to 2016 (to give a balanced panel). This led to a final sample of 241,908 patient encounters across 866 physician-procedure pairs (a total of 489 physicians). Table 1 presents summary statistics for the patient encounter level data set, as well as for the physician-procedure level data set.

In the next section, we test the effect of feedback valence (i.e., positive or negative feedback) on physician performance. We measure feedback valence in several ways and use two separate quasi-experimental research designs, specifically, a regression discontinuity design and two difference-in-difference designs, to test our theory. Our outcome of interest is always physician performance, which we measure using the patient's outcome: if the patient does not experience in-hospital mortality or 30-day readmission during the procedure, they are identified as having a "good outcome." Good outcomes thus allow us to measure physician performance. We use mortality and 30-day readmission because these are the metrics that physicians are given feedback on by the report card and because they are two of the most heavily validated and commonly accepted metrics for healthcare quality in the United States (Jencks et al. 2009).

In Section 3.1, a regression discontinuity design leveraging the report card's arbitrary performance zone cutoffs is used to evaluate feedback valence's effect on performance. Section 3.2 introduces two measures of feedback valence: "individual" and "relative" performance feedback. The former captures new performance information available to an individual surgeon but not to patients, whereas the latter captures physician performance relative to their peers. This section applies a difference-in-difference/event study design to assess feedback valence's impact on physician performance. Section 3.3 offers evidence to discount alternative explanations for the findings. (Note that we present many results from analyses in the form of both figures and tables. When not explicitly referenced in the main text, tables can be found by referring to the text notes at the bottom of the relevant figures.)

3.1. Empirics and Results: Regression Discontinuity Design

In this section, we use a unique design feature in the report card, specifically the zone "cutoffs," to test our hypotheses on feedback valence. For each physician-procedure, the report card complication rate falls into one of three zones: green (above average), yellow (average), and red (below average). These three zones are separated by two arbitrary cutoffs, the green-yellow cutoff and the yellow-red cutoff, chosen by the report card's designers based on national performance distributions and not based on any markers of clinical significance. Importantly, these cutoffs were unknown to physicians prior to the report card release. These features creates a distinct discontinuity in the treatment (feedback valence) at each cutoff, offering a solid basis for causal inference.

Our regression discontinuity design (RDD) compares physicians just below and above each cutoff. This analysis hinges on the idea that physicians near the zone cutoffs are similar in key aspects like quality and ability but differ mainly in the valence of the feedback provided to them. For instance, two physicians with

Table 1. Physician Analysis: Summary Statistics

Panel A

Variables	Mean	Standard deviation	Minimum	p25	p50	p75	Maximum
Age (yr)	68.57	8.75	51	62	69	75	115
Sex = female	0.57	0.49	0	0	1	1	1
Hispanic	0.02	0.15	0	0	0	0	1
Num Elixhauser comorbidities	2.06	1.52	0	1	2	3	12
Elixhauser mortality score	1.58	5.42	−16	0	0	3	56
Elixhauser readmission score	5.33	8.29	−3	0	0	8	91
Race = Black	0.06	0.24	0	0	0	0	1
Insurance = Medicare	0.70	0.46	0	0	1	1	1
Insurance = Private	0.26	0.44	0	0	0	1	1
Good Outcome	0.91	0.28	0	1	1	1	1
Observations	241,908						

Panel B

Variables	Mean	Standard deviation	Minimum	p25	p50	p75	Maximum
Green Zone	0.072	0.26	0	0	0	0	1
Yellow Zone	0.85	0.36	0	1	1	1	1
Red Zone	0.081	0.27	0	0	0	0	1
Report Card Complication Rate (%)	2.83	0.97	1.30	2.20	2.60	3.30	10.1
No. of competitors Within 25 miles	21.1	16.3	1	9	17	29	77
Positive feedback (Individual)	0.50	0.50	0	0	0	1	1
Positive feedback (Relative)	0.47	0.50	0	0	0	1	1
Physician experience (yr)	24.0	8.25	3	17	24	30	60
Physician sex = female	0.004	0.06	0	0	0	0	1
No. of patients	279.3	316.1	6	109	182.5	338	3,707
Observations	866						

Notes. Panel A presents summary statistics for the patient encounter level data set. Panel B presents summary statistics for the physician-procedure level data set. Good outcome = 1 if the patient does not experience in-hospital mortality or 30-day readmission during their surgery. Report Card Complication Rate is the adjusted complication rate provided by Propublica for each physician; it measures the % of surgeries performed by a physician (for a given procedure type) that result in in-hospital mortality or 30-day readmission. Individual and Relative positive feedback measures described in Section 3.2.

almost identical scores might receive different feedback valences: one barely making it to the green zone gets positive feedback, whereas the other, marginally falling into the yellow zone, gets less favorable feedback. Comparing the changes in performance of these closely ranked physicians isolates the effect of feedback valence on performance. However, as physicians move further away from the cutoff, their reactions to report card valence might start reflecting unobservable differences between them, potentially skewing our valence effect estimates.

We conduct our regression discontinuity analyses at the physician-procedure level, performing two separate analyses for the green-yellow and yellow-red cutoffs. The key independent variable in each analysis is the physician's proximity to either the green-yellow or yellow-red cutoff, measured in the specific scale assigned for that physician and procedure.[8] Our outcome of interest is unadjusted physician performance in the post–report card period (i.e., proportion of surgeries with good outcomes).

We estimate both nonparametric and parametric RDD models, although we present results for the latter only in the Online Appendix. Table 2 presents estimates from nonparametric robust local polynomial inference methods, using optimal covariate-adjusted bandwidths equally set around the cutoff and robust bias-corrected standard errors (via the *rdrobust* package; Calonico et al. 2017). The "baseline model" incorporates procedure dummies and preperiod unadjusted good outcome rates as covariates. Additional models are estimated including preperiod patient characteristics (the "Pre-Patchars" model) and excluding any covariates (the "NoCov" model). We present estimates from both heteroskedasticity-robust standard errors and errors clustered by procedure in the baseline model. We first present results for the green-yellow cutoff before analyzing the yellow-red cutoff.

In line with our hypothesis, results show that receiving more positive feedback led to an increase in physician performance. Specifically, physicians just left of the green-yellow cutoff (in the green zone) exhibited stronger performance for a given procedure postperiod than those just right of the cutoff, regardless of covariates or clustering method (Table 2, columns (1) to (4)).

Table 2. Physician Analysis: Nonparametric Regression Discontinuity Estimates

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate	−0.052*	−0.070*	−0.095***	−0.045	0.230	0.002	0.005	−0.003
	(0.0335)	(0.0413)	(0.0345)	(0.0334)	(0.241)	(0.0107)	(0.0220)	(0.0214)
Observations	864	864	864	865	864	864	864	864
Model	Baseline	Pre patchars	Pre patchars	NoCov	Baseline	Baseline	Baseline	Pre patchars
Kernel	Triangular	Triangular	Triangular	Triangular	Triangular	Triangular	Triangular	Triangular
Mean left	0.928	0.928	0.928	0.928	0.928	0.923	0.918	0.918
Mean right	0.915	0.915	0.915	0.915	0.915	0.910	0.900	0.900
Robust p value	0.097	0.066	<0.001	0.139	0.457	0.884	0.798	0.967
Regression bandwidth	1.84	1.60	1.43	2.03	1.46	4.62	4.38	4.24
Effective no. of observations (L)	50	48	48	50	35	315	96	96
Effective no. of observations (R)	93	79	57	93	12	253	57	57
Bias bandwidth	4.02	3.56	3.63	3.94	4.19	6.53	7.21	7.16
Order polynomial	1	1	1	1	1	1	1	1
Standard errors clustered by procedure	N	N	Y	N	N	N	N	N

Notes. Dependent variable is *Physician performance in the Post Period*. Unit of analysis is the physician-procedure. The outcome variables for all columns is physician performance (good outcomes, i.e., proportion of encounters without in-hospital mortality or 30-day readmission) in the post–report card period. Columns (1) through (4) estimate the nonparametric effect of the discontinuity at the green-yellow cutoff, whereas columns (7) and (8) estimate it for the yellow-red cutoff. Columns (5) and (6) estimate "placebo" effects on the left and right of the green-yellow, respectively. The "baseline" model (columns (1), (5), (6), and (7)) include dummies for procedure and pre–report card physician performance as covariates. The "Pre pat chars" (columns (2), (3), and (8)) model adds in pre–report card patient characteristics (age, sex, race, insurance status, Elixhauser comorbidities, and ethnicity) as covariates to the baseline model. The "NoCov" model (column (4)) has no covariates. Heteroskedasticity-robust standard errors in parentheses for all models, but those for column 3 are clustered at the level of the procedure. Only robust bias-corrected p values are reported.

***$p < 0.001$; **$p < 0.05$; *$p < 0.1$.

RDD estimates indicate a significant performance increase of 4.5–9.5 percentage points, translating to a 4.9%–10.3% improvement on the 91.5% sample mean of unadjusted good outcomes for those left of the cutoff. Notably, RDD estimates without covariates are similar to those with covariates but have slightly more imprecise standard errors (column (4), $p = 0.13$), supporting their inclusion for increased statistical precision (Cunningham 2021).

Figure A.2 in the Online Appendix displays these results graphically by showing the nonparametric RDD plots with varying distances around the green-yellow cutoff (2, 2.5, or 3 units), different covariates (procedure or physician fixed effects), and polynomial fits (1 or 2). Negative distance values on the x axis represent the green zone, and positive values represent the yellow zone. Consistent with our hypothesis, surgeons just within the green zone (receiving more positive feedback) show more improvement than those just outside it (receiving less positive feedback). Across all specifications, a clear discontinuous drop in postreport card performance is evident when transitioning from left (green zone) to right (yellow zone) of the cutoff, even after adjusting for the surgeon's prereport card performance. This effect is more pronounced when narrowing the bandwidth closer to the cutoff, with the tradeoff being that data are sparser near the cutoff.

In the Online Appendix, Section A.1, we discuss results from the parametric models, which echo the nonparametric results. We also perform several tests to check the validity of our findings around the green-yellow cutoff. Theoretically, the continuity assumption (which posits no sudden changes at zone cutoffs in the absence of treatment) is almost certainly valid in our setting because, given physicians' lack of prior knowledge about the report card and cutoffs, outcome manipulation or gaming of the discontinuity is unlikely. However, in the Online Appendix, we present histograms of the run variable to demonstrate no nonrandom heaping at the cutoff, results from the McCrary density test, graphs and estimates from covariate balance tests, and estimates from "placebo" cutoffs as recommended by Imbens and Lemieux (2008). None of these tests endanger our main RDD estimates. These approaches collectively bolster confidence in our findings, demonstrating a robust and valid green-yellow RDD analysis.

Conversely, the RDD analysis surrounding the yellow-red cutoff suggests that receiving more negative feedback does not significantly affect physician performance (columns (7) and (8) in Table 2).[9] In other words, physicians just to the right of the yellow-red cutoff (i.e., receiving negative feedback) did not perform significantly worse than physician just to the left of it (i.e., receiving less negative feedback).

Overall, results from RDD analyses of both cutoffs are consistent with the idea that positive feedback is more beneficial than negative feedback. Finding significant effects around the green-yellow cutoff (positive versus less positive feedback) but not around the yellow-red cutoff (less negative versus negative feedback) suggests

that positive feedback leads to improvements in physician performance but negative feedback may not be detrimental. However, later analyses do provide some support for detrimental effects of negative effects. We bolster our conclusions with a difference-in-difference approach in the next section, aiming to tackle our research question from various perspectives. This strategy ensures that the limitations of any one method are compensated by the strengths of the others.

3.2. Empirics and Results: Difference-in-Difference Design

For this quasi-experimental design, we use a 2×2 difference-in-difference (DiD) design (positive/negative feedback × pre/post report card). The design involves a one-time treatment with a balanced panel, which mitigates many issues highlighted in recent DiD literature (Callaway and Sant'Anna 2021, Goodman-Bacon 2021, Sun and Abraham 2021). The core assumption in the DiD design is parallel trends in performance: physicians receiving positive or negative feedback should exhibit similar performance trends prior to the report card. With parallel trends, any performance changes after the report card release (i.e., after quarter 14) can be attributed to the report card. We support this assumption through various analyses and graphical evidence.

We create two measures of positive versus negative feedback: (i) individual performance feedback (information in the report card pertaining to the individual physician's performance), and (ii) relative performance feedback (information in the report card pertaining to the physician's performance relative to their peers).

3.2.1. Individual Performance Feedback. This measure identifies positive and negative feedback by comparing a physician's perceived performance, inferred from their own raw observed good outcomes prior to the report card, against report card feedback. The report card provides new information to physicians about their own performance because it adjusts for uncontrollable factors more completely than physicians can themselves. Physicians cannot be expected to appropriately adjust due to their inability to compare their own outcomes to others' and the small, nonrepresentative nature of their patient samples (e.g., averaging 25 surgeries per quarter in our data). Thus, we use raw observable outcomes as a noisy but unbiased proxy for each physician's priors over their own performance (although we test the strength of this assumption in two robustness checks later). Negative feedback is identified when report card feedback is worse than these observed outcomes, and positive when it is better. For example, a surgeon with a 6% raw complication rate for hip surgeries, but a 4% adjusted report card complication rate, is considered to have received positive feedback.

This method draws from the study of Kolstad (2013) on report cards' impact on physician performance.[10] It reflects extensive research in economics, medicine, and psychology highlighting physicians' struggles with risk adjustment, prediction, and probability estimation due to heuristic biases, medical complexity, limited patient samples, and lack of statistical training (Christensen-Szalanski and Bushyhead 1981, Elstein et al. 1981, Eddy 1982, McNeil et al. 1982, Redelmeier and Tversky 1990, Redelmeier and Shafir 1995, Croskerry 2003, Gigerenzer et al. 2007, Berner and Graber 2008). Therefore, their performance perceptions are likely formed from personal observations, not in-depth statistical analysis.

3.2.2. Relative Performance Feedback. With this measure, we categorize feedback valence based on report card information about a physician's peers. Recall that each physician is categorized into red (below average), yellow (average), or green (above average) zones for each procedure performed. The physicians also see the placement within these zones of all their peers within 25 miles. Relative performance feedback valence is determined by assessing the physician's zone relative to their peers', which informs us about whether a physician noticeably outperforms or underperforms most of their peers. Following Zimmermann (2020), feedback is classified as positive when physicians have more peers in worse zones and negative when they have more peers in better zones. For instance, a physician in the yellow zone for a hip procedure receives negative feedback if 50% of peers are in green zones and 20% in red, indicating performance below peers.[11]

Our aim in constructing this measure is to match experimental research that manipulates valence by exogenously varying the performance of an individual's peer group (i.e., randomly selecting peers), holding constant individual performance (Eil and Rao 2011, Zimmermann 2020, Erickson et al. 2022). To this point, our setting offers plausibly exogenous variation in the peer group (independent of a surgeons' own performance) due to (i) the arbitrary cutoffs used to delineate the green/yellow/red zones for each procedure and (ii) the arbitrary choice to compare physicians to peers within 25 miles of their primary hospital. Both these choices, the cutoffs and the radius of 25 miles, provide variation in the peer group that, from the physicians' perspective, could not be anticipated and is effectively random. This variation is especially salient if we exclusively focus on physicians in the yellow zone, in which case the measure of valence depends solely on variation in peer performance.[12]

Using these two measures of feedback valence (individual and relative, first separately and then together), we apply the following linear probability model on a patient encounter-level data set, which indexes each patient encounter by i, where one of the eight key

procedures s was performed by physician j at hospital h during quarter q:

$$Y_{i(sjhq)} = D_{js} \times \sum_{\substack{q=1 \\ q \neq 14}}^{20} \boldsymbol{\beta_q} I(q - q^* = 15) + \mathbf{REPCARD}_{js} + \mathbf{X_i} + \boldsymbol{\beta_s} + \boldsymbol{\beta}_{\boldsymbol{ZIP}_i} + \boldsymbol{\beta}_{\boldsymbol{HOSP}_i} + \varepsilon_{i(sjhq)}, \quad (1)$$

where

- $Y_{i(sjhq)}$ is equal to one if patient encounter i had a good outcome (i.e., did not end in in-hospital mortality or 30-day readmission);
- $D_{js} \times \sum_{\substack{q=1 \\ q \neq 14}}^{20} \beta_q I(q - q^* = 15)$ represents a fully saturated interaction[13] between:
  - An indicator of the treatment i.e., whether physician j received positive feedback from the report card for procedure s (using individual OR relative feedback)
  - Indicators for every quarter from 2012 through 2016 (with quarter 14 as the omitted category), measuring time since the report card release in quarter 15.
- $REPCARD_{js}$ is a vector of report card features: the report card complication rate, zone, and number of peers for physician j's performing procedure s[14];
- X_i is a vector of patient characteristics specifically, age, sex, ethnicity, race, insurance status, Elixhauser mortality and readmission indices (Elixhauser et al. 1998, Moore et al. 2017). Elixhauser indices are the most commonly used and validated measures of patient comorbidity based on 29 different, critical comorbidities;
- X_j is a vector of physician characteristics, specifically, sex and years of prior experience;
- β_s is fixed effects for procedure s;
- $\beta_{ZIP\ i}$ is fixed effects for the zip code of the patient's residence; and
- $\beta_{HOSP\ i}$ is fixed effects for the hospital h.

We present estimates using heteroskedasticity-robust standard errors, clustering at the physician level (Wooldridge 2010, Mullahy and Norton 2022). The β_q coefficient measures the effect of receiving positive (versus negative) feedback on physician performance at each quarter, relative to the baseline performance in quarter 14. In accordance with current DiD methodologies (Miller et al. 2021) and best practices in empirical economics (Cunningham 2021), we plot treatment lead and lag regression coefficients using an event study design, followed by presenting DiD regression results in tabular form (for which we replace the quarter indicators with a postreport card indicator, $Post_{iq}$).

However, before decomposing performance by feedback valence, we first show how physician performance changes *on average* from pre– to post–report card release at each quarter. Figure A.5, (a) and (b), of the Online Appendix, shows a positive trend in both unadjusted and adjusted (using Equation (1)) rates of good outcomes, respectively. Overall, physician performance improves by 1.46 percentage points on average post–report card release (Figure A.5(c) of the Online Appendix). However, this improvement may not be solely attributable to the report card due to a preexisting positive time trend in good outcomes prior to the report card release. This pretrend challenges the validity of a simple prepost design for causal inference.

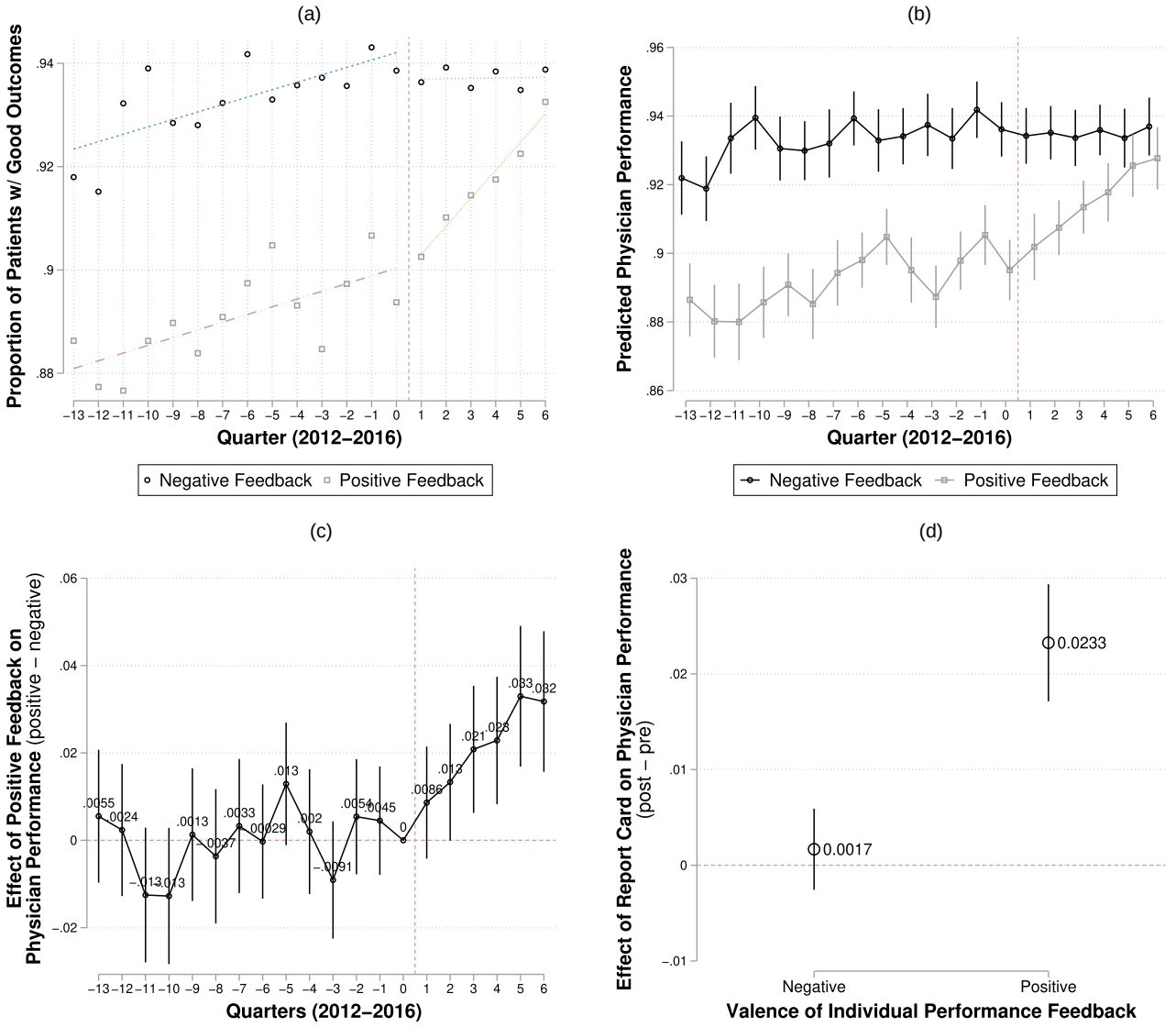
To overcome this limitation, we compare physicians receiving positive versus negative feedback. Unlike the overall trend, the performance difference between these two groups remained stable before the report card release. This allows us to use a DiD approach, adhering to the parallel trends assumption, thereby enabling a more robust estimation of the report card's causal effect on physician performance.

The DiD graphs are presented in Figure 2 for the individual feedback valence measure and in Figure 3 for the relative feedback valence measure. The DiD estimates are presented in Table A.3 of the Online Appendix. Panel (a) in both figures present a preliminary unadjusted analysis aimed at bolstering confidence in our subsequent adjusted analyses. Both figures show that, before the report card release, the unadjusted proportions of good outcomes trended similarly between physicians receiving positive and negative feedback, regardless of which measure we use for feedback valence. Yet physicians who received positive feedback from the report card exhibit significantly improved performance after its release compared with their counterparts who received negative feedback.[15]

Panels (b) and (c) in both figures show that these unadjusted results hold even after adjusting for a rich set of covariates (Equation (1)). Specifically, prior to the report card, physicians with both positive and negative feedback exhibited similar good outcome trends. Postrelease, however, only those with positive feedback showed a marked improvement in performance.

Panel (d) in both figures present the report card's marginal effects on physicians receiving positive and negative feedback. Figure 2(d) reveals that physicians with positive feedback showed a 2.3-percentage-point (pp) increase in performance, whereas negative feedback led to no significant change, translating to a 2.1-pp DiD estimate (Table A.3, column (1), in the Online Appendix) or a 2.5% rise from the 90.8% prereport average. Likewise, Figure 3(d) indicates that surgeons with positive feedback improved by 1.9 pp, notably more than those with negative feedback, who improved by less than 1 pp. This corresponds to a 1-pp DiD estimate and a 1.1% improvement from the preperiod mean (Table A.3, column (4), in the Online Appendix).[16] Interestingly, these

Figure 2. (Color online) Physician Analysis: (Event Study) Effect of Feedback Valence (Using Individual Performance Feedback) on Physician Performance

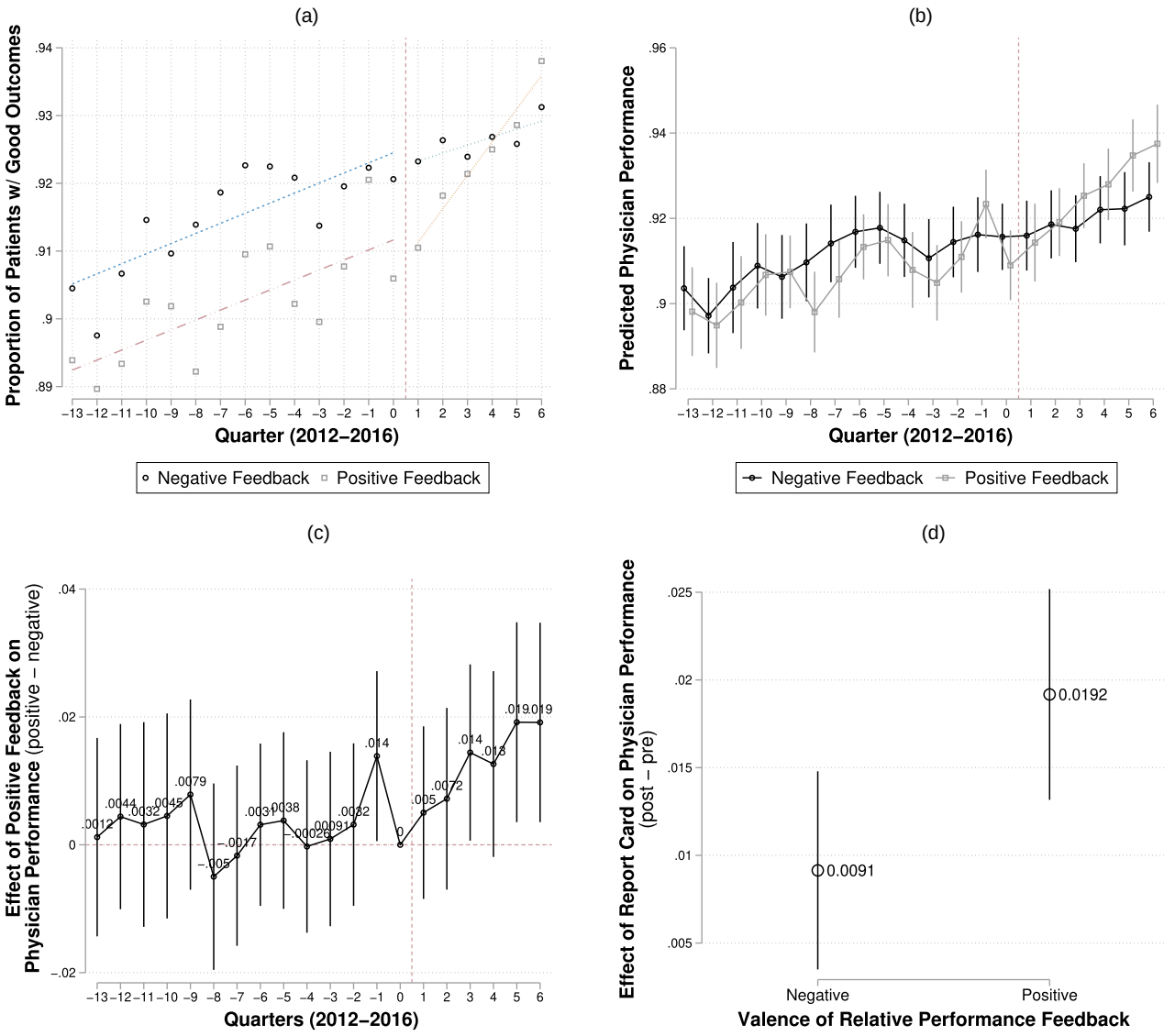


Notes. (a) Unadjusted performance (raw means). (b) Adjusted performance (fitted values). (c) DiD/event study graph. (d) Marginal effects. Unit of analysis is the patient encounter. Negative and positive feedback are defined using individual performance feedback, as explained in Section 3.2. The vertical dotted line represents the time at which the report card was released. Panel (a) plots the unadjusted physician performance (raw proportion of good outcomes, i.e., encounters that did not experience in-hospital mortality or 30-day readmission) aggregated at each quarter from 2012 to 2016. Panel (b) plots the fitted values for physician performance from Equation (1) at each quarter separately for physicians receiving negative and positive feedback (at the means of all other covariates). Panel (c) plots the effect of positive feedback from Equation (1) at each quarter, with quarter 0 as the omitted category. Panel (d) plots the marginal effect of the report card on physician performance from Equation (1), separately for physicians receiving positive and negative feedback. Heteroskedasticity-robust standard errors are clustered at the level of the physician. 95% confidence intervals are presented for each estimate.

effect sizes for the RDD in the previous section are larger (although not implausibly so) than these DiD estimates. The larger magnitude suggests that unobservables (such as physician quality and patient pool), which the RDD estimate should account for, may be biasing our DiD estimates toward the null.

3.2.3. Specification-Based Robustness Checks. First, we attempt to determine the extent to which these feedback-valence effects are driven by the beneficial effects of positive feedback versus the detrimental effects of negative feedback. Referring to surgeons with positive feedback as the treatment group and those with negative feedback as the control group is misleading because both are affected by the report card. Thus, as a robustness check, we use performance on *nonkey procedures* as a control group, as these surgeries are not evaluated by the report card and should show smaller

Figure 3. (Color online) Physician Analysis: (Event Study) Effect of Feedback Valence (Using Relative Performance Feedback) on Physician Performance



Notes. (a) Unadjusted performance (raw means). (b) Adjusted performance (fitted values). (c) DiD/event study graph. (d) Marginal effects. Unit of analysis is the patient encounter. Negative and positive feedback are defined using relative performance feedback, as explained in Section 3.2. The vertical dotted line represents the time at which the report card was released. Panel (a) plots the unadjusted physician performance (raw proportion of good outcomes, i.e., encounters that did not experience in-hospital mortality or 30-day readmission) aggregated at each quarter from 2012–2016. Panel (b) plots the fitted values of physician performance from Equation (1) at each quarter separately for physicians receiving negative and positive feedback (at the means of all other covariates). Panel (c) plots the effect of positive feedback from Equation (1) at each quarter, with quarter 0 as the omitted category. Panel (d) plots the marginal effect of the report card on physician performance from Equation (1), separately for physicians receiving positive and negative feedback. Heteroskedasticity-robust standard errors are clustered at the level of the physician. 95% confidence intervals are presented for each estimate.

(if any) postquarter 14 performance changes. Importantly, we apply physician fixed effects to compare each physician's performance on their key versus nonkey procedures. In our data, approximately 70% of surgeries are key procedures, reflecting surgeons' specialization. Results in Figure A.6 in the Online Appendix show that as expected, performance on control surgeries does not change in the "post" period. Moreover, we observe that even within-physician, positive feedback consistently boosts performance ($p < 0.05$ with both individual and relative feedback), whereas negative feedback may adversely affect performance (marginally significant with individual feedback ($p = 0.085$) but not with relative feedback ($p = 0.840$), although the effect is negative signed as hypothesized).

Second, we verify that individual and relative feedback valences are distinct and independently impact physician performance. By estimating a DiD equation

that includes interactions between the $Post_q$ variable and *each* type of feedback valence measures simultaneously, we establish their marginal effects. Table A.3, columns 7–9, in the Online Appendix, illustrates that interactions of $Post_q$ with both individual and relative feedback valence measures do not alter our conclusions. Column (7) indicates positive individual feedback improves performance by 2 pp more than negative individual feedback, and positive relative feedback by 0.7 pp more than negative, with each feedback type independently controlled. This suggests the feedback dimensions convey unique information and affect surgeon performance through different pathways. These results remain robust after adding year and quarter fixed effects (column (8)) and physician fixed effects (column (9)).

Third, we test the robustness of our DiD estimates to the length of the pretreatment window. Thus far, we use the longest pretreatment period available (14 quarters before report card release) to support the parallel trends assumption in our DiD analysis, even though the postperiod quarter length is six quarters. However, literature on the optimal pretreatment window length for DiD is unsettled. To our knowledge, only two papers obliquely mention this issue: Chabe-Ferret (2015) finds that symmetric DiD estimators (equal pre- and posttreatment periods) outperform matching and are consistent under certain conditions, whereas Slaughter (2001) uses rolling pretreatment windows for robustness. Based on these, we demonstrate that omitting quarters sequentially does not alter our DiD estimates (Table A.5 in the Online Appendix). That is, restricting the number of pretreatment quarters from 13 (i.e., dropping preperiod quarter −13, such that there are 13 preperiod quarters) all the way down to 5 (i.e., dropping preperiod quarters −13 through −5, such that there are 5 preperiod quarters) does not impact the results, either with the individual measure (Table A.5, top panel) or the relative measure (Table A.5, bottom panel).

Finally, we perform two additional robustness checks to allow for physicians forming priors over their performance in slightly more realistic ways with our measure of individual feedback valence: (i) by accounting for physicians' potential adjustments for patient age and sex when forming priors,[17] and (ii) considering a 50% error rate in their self-assessment.[18] Our results remain robust to either measure (Figure A.8, (a) and (b), in the Online Appendix), confirming that our results are not overly reliant on raw outcome measures or perfect self-evaluation by physicians.

3.2.4. Heterogeneities by Surgeon's Preperiod Performance. We next examine whether the effects of feedback valence depend on average performance in the preperiod, which is measured as average (within-procedure) patient outcomes in the preperiod adjusted for patient characteristics and hospital quality. This analysis allows us to further isolate the effects of feedback valence from mechanistic trends arising from preexisting differences in performance levels between physicians receiving positive versus negative feedback (Panel (a) in Figures 2 and 3).

Our first analysis amends the DiD equation by adding the continuous measure of preperiod performance ($Perf^c_{js}$) as an additional covariate and interacting it with $Post_{iq}$.[19] If the effects of feedback valence are driven by differences in preperiod performance levels, then including the $Post_{iq} \times Perf^c_{js}$ should eliminate the effects of $Post_{iq} \times$ valence ($D_{js} = 1$ for positive feedback). However, we find that physicians still improve more after receiving positive rather than negative feedback, even after accounting for preperiod performance. Specifically, although higher preperiod performance is always associated with a decrease in postperiod performance (i.e., the coefficient on $Post_{iq} \times Perf^c_{js}$ is always significantly negative, likely due to mean reversion), the coefficient on $Post_{iq} \times D_{js}$ remains significantly positive (untabulated: $\beta = 1.05$ percentage points, $p < 0.001$ for the individual valence measure, and $\beta = 1.07$ percentage points, $p = 0.003$ for the relative valence measure).

To expand on this analysis, we categorize physicians into high and low performer types by taking a median split of their adjusted performance in the preperiod. We then amend Equation (1) by adding this binary variable (referred to as $Perf_{js}$) and estimating the full triple interaction model. Results in Figure A.7, (a)–(d), in the Online Appendix, show that, with both our individual and relative feedback measures, positive feedback leads to greater performance improvements for both low and high performer type physicians than negative feedback.[20] In fact, the effect of feedback valence is nearly identical for both physician types. Thus, the previously discussed effects of valence do not appear to be related to any differences in average preperiod performance (which may exist due to differences in physician ability, which we cannot observe but proxy for by using performance types in this analysis). Interestingly, examining marginal effects of the report card (using either feedback measure) suggests that, although both low and high performer type physicians receiving positive feedback improve, high performer types who receive negative feedback experience a subsequent deterioration in their performance.

3.3. Theory-Based Robustness Checks and Alternative Explanations

We next discuss how we rule out several alternative explanations, before outlining the controlled experiments conducted to replicate and extend our key findings in the next section.

3.3.1. Physician Financial Incentives. We propose that our findings—physicians responding more to positive than negative feedback—are rooted in behavioral factors. Yet, it might also be a response to financial incentives. Stronger financial incentives from the report card release, favoring positive over negative feedback, could drive this pattern. We evaluate this possibility using two measures: competitors within 25 miles of physician j for procedure s, and the share of j's patients with private insurance prereport card. Given that both competition and higher private insurance reimbursements incentivize improvement, physicians under more of these pressures are expected to enhance performance postreport card. Our regression analysis, based on a period (t), physician (j), and procedure (s) level data set, where period t is 1 postreport card and 0 otherwise, tests this hypothesis using the following equation:

$$Y_{jst} = Post_t \cdot D_j \cdot X_{js} + \mathbf{REPCARD}_{js} + \chi_{\mathbf{jst}} + \boldsymbol{\beta_s} + \varepsilon_{jst}, \tag{2}$$

where

- $Y_{(jst)}$ is the outcome of interest, such as the (mean) unadjusted physician performance (proportion of patients with Good Outcome) by physician j for procedure s in period t
- $Post_t = 1$ if representing the post–report card period and 0 otherwise
- D_j is equal to one if the physician receives positive performance feedback (individual or relative)
- X_{js} is one of two measures of financial incentives: number of competitors faced by physician j for procedure s, and proportion of physician j's procedure s patients who paid using private insurance in the preperiod
- $\mathbf{REPCARD}_{jq}$ is report card complication rate and zone
- χ_{jst} is a vector of average patient characteristics: age, sex, race, ethnicity, Elixhauser comorbidities, and insurance type
- β_s is fixed effects for procedure

Table A.6 (columns (1)–(4), Panels A, B, and C), of the Online Appendix, indicates financial incentives do not alter feedback valence effects. The triple interaction coefficient is insignificant (Panel A), and the DiD estimates of report card valence (using individual or relative performance feedback) are consistent and significant across the 25th, 50th, and 75th percentiles of financial incentive measures (number of competitors or patient private insurance proportion). Thus, positive feedback enhances physician performance more than negative feedback across various financial incentive levels. This aligns with Kolstad (2013), demonstrating that nonfinancial (behavioral) incentives from report cards outweigh financial incentives in influencing physician performance.

3.3.2. Patient Demand Effects. We also explore possible demand effects, that is, patients selecting physicians based on report card data. Yet, these effects are unlikely to account for our findings for several reasons.

First, extensive literature suggests that public data on provider quality (e.g., report cards) scarcely affects consumer choices due to factors like unawareness, insurance constraints, or difficulty interpreting data (Uhrig and Short 2002, Hirth et al. 2003, Epstein 2010, Grabowski and Town 2011, Jung et al. 2011, Scanlon et al. 2015). When effects do occur, they vary by outcomes reported (Mukamel et al. 2008), patient types (Jin and Sorensen 2006), provider types (Clement et al. 2012), and market conditions (Chou et al. 2014). Notably, Shi et al. (2017) found in a national study that only 12% of patients knew about physician report cards initially, a figure consistent with other studies (Schneider and Epstein 1998, Abraham et al. 2004, Sinaiko et al. 2012, Christianson et al. 2014, Scanlon et al. 2015). They also observed that a new report card only raised awareness by 3.8 pp among the unaware and only if it was relevant to their community and chronic condition. Given this scenario where consumers are generally uninformed or unaffected by physician report cards, demand effects likely do not play a significant role in our setting.

Second, regarding the feedback valence measure using individual performance feedback, demand effects are not an issue as patients do not see the surgeon's raw complication rate. Consequently, they cannot select into treated physicians based on this measure (those receiving positive individual performance feedback).

Third, if patients act on report card data, we would expect higher patient volumes for green zone physicians and lower for red zone ones, as zones are clear feedback valence indicators. However, Table A.6, column (5) (Panels A and D), in the Online Appendix, reveals the contrary: Green zone physicians experience the most significant reductions in patient volumes among the three zones.

Fourth and finally, for patient demand to skew our results, healthier patients would need to select physicians with positive feedback. Contrarily, Table A.6, column (5) (Panels A and D) shows the opposite: Only yellow zone physicians see a (nonsignificant) drop in patient comorbidity (a measure of medical complexity/severity), whereas green and red zone physicians experience a (nonsignificant) increase in comorbidity. Thus, patient demand likely does not drive our results.

3.3.3. Mean Reversion. Mean reversion is a typical concern in feedback contexts, but it is not problematic in our study for several reasons. First, the report card is designed to account for mean reversion by adjusting

top and bottom physicians' performances. Second, mean reversion would imply that physicians with positive feedback worsen, and those with negative feedback improve, which is contrary to our findings. Last, our methodologies—DiD and regression discontinuity (RDD)—guard against mean reversion. In DiD, if mean reversion influenced postperiod outcomes, distinct pre-event trends between physicians with positive and negative feedback would be evident. In RDD, mean reversion is not a factor as it compares physicians with similar quality and performance levels.

3.3.4. Cherry Picking. The possibility of physicians selectively choosing patients ("cherry picking") after the report card could bias our results, if those receiving positive feedback prefer healthier patients, whereas those with negative feedback opt for, or end up with, sicker ones. However, Table A.6, column (6) (Panels A and D), in the Online Appendix, indicates this is not an issue. In fact, the trend is contrary: Physicians in both green and red zones are treating more comorbid patients, and those in the yellow zone are treating fewer comorbid patients than before.

4. Experimental Analysis

4.1. Setting and Data

To augment our findings from the physician analysis, we conducted four controlled experiments with 1,001 U.S. participants using Prolific for experiment 1 and Amazon's Mechanical Turk for experiments 2a, 2b, and 2c.[21] All experiments were approved by the institutional review board. These experiments allow us to (i) definitely identify causal effects, (ii) precisely measure individual performance while ruling out factors like cherry picking and demand effects, and (iii) assess moderators. Experiment 1 uses relative performance feedback, whereas 2a, 2b, and 2c use individual performance feedback. Experiments 1 and 2a focus on isolating the effects of feedback valence; 2b and 2c examine factors aimed at addressing issues with negative feedback.

In all experiments, participants play a multiarmed bandit game in which they choose among three buttons on each trial (100 trials in experiment 1 and 140 trials in experiments 2a, 2b, and 2c). Each button elicits "points" equal to its true value plus a uniformly distributed mean-zero error term. Participants were paid based on the total points they earned across all trials. The more quickly they learned which buttons had higher true values, the more points they could earn. Their understanding of the game was evident as they earned more points than they would have from choosing randomly ($p < 0.001$ for each experiment).[22]

The multiarmed bandit game is widely used in studies on learning from feedback (Posen and Levinthal 2012, Christian and Griffiths 2016, Lefebvre et al. 2017). This task requires participants to (i) efficiently process information by updating beliefs about the button's value after each press and (ii) implement effective strategies, balancing exploring new buttons with exploiting known ones. Figure A.9 in the Online Appendix gives a screenshot of the task, and Online Appendix B.1 offers more details on the task, procedures, and incentives.

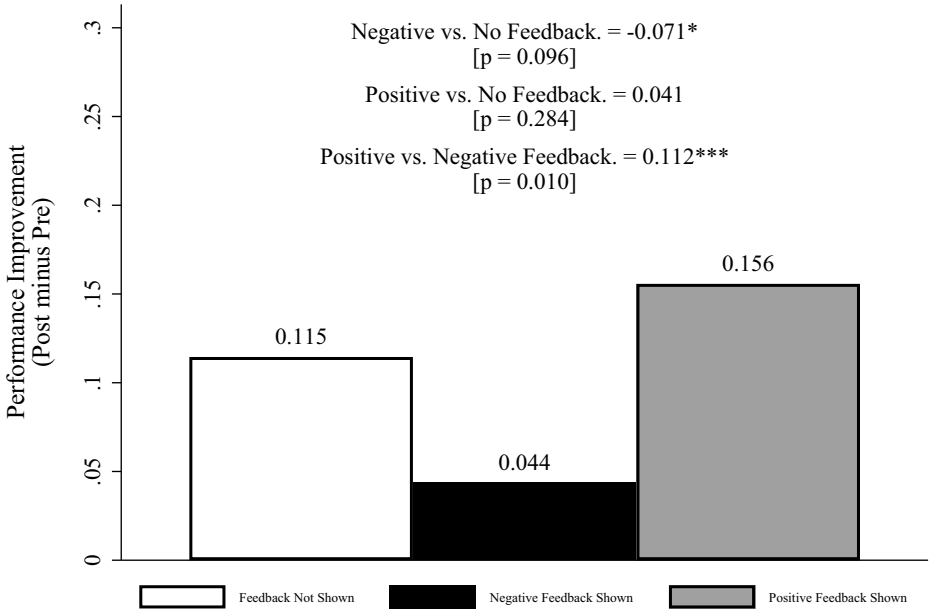
Three aspects of our experiments stand out. First, we assign feedback valence exogenously (without using deception), isolating it from traits like ability. Second, we use control conditions without feedback, providing a clean counterfactual. Third, the feedback manipulations always occur during a surprise midtask break. Thus, all conditions are identical in the first half.

4.1.1. Experiment 1: Design. Experiment 1 uses a 1×3 between-subjects design with three feedback conditions during the surprise break halfway through (after trial 50): no feedback, negative feedback, and positive feedback. During the break, those in the control group, no feedback shown, are only reminded they are halfway done. Those in the treatment groups see their first-half performance (total points) compared with five participants who previously completed the task. Feedback valence is manipulated by randomly selecting the peer group (more details below) (Eil and Rao 2011, Coutts 2019, Zimmermann 2020, Erickson et al. 2022). The feedback presentation format is similar to the ProPublica report card, placing participants on a horizontal, color-coded scale. Low/average/high performers are categorized into red/yellow/green zones, where the red/green zones consist of those in the bottom/top quintile of performance. In the Online Appendix, Figure 1 illustrates the ProPublica report card, and Figure A.10 displays the experiment's feedback.

Feedback valence is determined by the performance of the peers to whom participants are compared. We collected data for a sample of peers before our main task and then chose 18 from this sample (six in each of the three zones). The peer group for each participant in the main experiment is then formed by taking an i.i.d. random sample (without replacement) of five peers from the pool of 18. We classify feedback as negative when participants have more peers in the green zone (top quintile) than in the red zone (bottom quintile), and feedback as positive otherwise. On average, participants will compare less favorably when they are assigned to more peers in the green zone and fewer peers in the red zone. Importantly, this manipulation of valence is based on random variation in peer assignment independent of participants' own performance.

4.1.2. Experiment 1: Results. Table A.7 shows descriptive statistics, and Figure A.11 shows that, as expected, performance is nearly identical across conditions in the preperiod. To formally examine how feedback affects

Figure 4. Experiment 1: Performance Improvement for Each Feedback Condition



Notes. This bar chart shows the average performance improvement from the pre to the post period in each of the three experimental conditions in experiment 1. The values are calculated from Equation (3) and the experimental conditions are described in Section 4.1.1.

performance, we estimate the following regression with heteroskedasticity-robust standard errors clustered by participant:

$$Performance_{ij} = Post_i \cdot FeedbackCondition_j + \varepsilon_{ij}, \qquad (3)$$

where

- $Performance_{ij}$ is participant j's performance on trial i, ranging from zero (choosing the worst button) to two (choosing the best button)
- $Post_i$ is one for trials 51–100 (post feedback break), and zero for trials 1–50
- $FeedbackCondition_j$ is zero for the control condition (no feedback shown), one for the treatment condition with negative feedback, and two for the treatment condition with positive feedback

Figure 4 shows the main results by giving the marginal effect of $Post_i$, that is, the participant's improvement from before to after the feedback break for each of the three feedback conditions.[23] Performance significantly improves from the first to second half when participants do not receive feedback and when they receive positive feedback, but not when they receive negative feedback. Moreover, negative feedback significantly reduces the performance improvement compared with the control condition with no feedback, as well as compared with the positive feedback condition. In contrast, positive feedback (insignificantly) improves the performance improvement compared with the control condition.

Additional analyses show that results are concentrated among competitive individuals. We classified participants as competitive if they somewhat or strongly agreed (on a five-point Likert scale) with the following statement from the postexperimental questionnaire: "I am a very competitive person." The differential response to negative and positive feedback was significant for competitive individuals but not for noncompetitive individuals. Compared with the control condition with no feedback, competitive individuals improve significantly less when they receive negative feedback ($p = 0.048$) and (insignificantly) more when they receive positive feedback ($p = 0.196$). However, there is no effect of negative feedback ($p = 0.623$) or positive feedback ($p = 0.874$) for noncompetitive individuals. Thus, more competitive individuals do not respond more effectively to both types of feedback. Rather, they respond even worse to negative feedback and even better to positive feedback.

Finding that the effects are concentrated among competitive individuals is particularly noteworthy because physician competitiveness stemming from pride and social comparison is a key explanation for why they respond to new performance information (Garcia et al. 2013, Kolstad 2013, Liao et al. 2016, Meeker et al. 2016, Song et al. 2018). Furthermore, physicians are likely more competitive than the average person given the selectivity of medical schools, residencies, and fellowships (Moon 2021); orthopedic surgery, the field we study, is perhaps the most selective (Murphy 2018).

Overall, experiment 1 corroborates our hypothesis that positive feedback is more effective than negative feedback. Finding consistent results with surgeons'

making consequential real-world decisions and laypersons in an abstract task suggests that the results are due to inherent behavioral factors (Lefebvre et al. 2017, Eskreis-Winkler and Fishbach 2019). There was one notable difference: Whereas the surgeon analysis found stronger evidence for positive feedback improving performance than for negative feedback harming performance, the experiment found somewhat stronger evidence for negative feedback harming than for positive feedback helping performance. However, we interpret this difference cautiously because the experimental results directionally support both positive feedback helping ($p = 0.284$) and negative feedback harming ($p = 0.096$), and the surgeon analysis found some support for both effects as well. Furthermore, across all analyses thus far, the directional effects point to both positive helping and negative harming.

There may also be various reasons for the difference. For example, feedback in the experiment is objective and participants cannot easily discount its quality. In contrast, physicians widely criticize the quality of report cards, such that those who receive negative feedback will easily dismiss it as inaccurate, whereas those who receive positive feedback readily accept it as valid (Ruzzene and Noller 1986, Gnepp et al. 2020). In the experiment, where participants cannot easily dismiss negative feedback, they may become demotivated by it (Deci et al. 1999, Hattie and Timperley 2007, Erickson et al. 2022) or they may be motivated to improve but fail to adapt their task strategies effectively (Hannan et al. 2008).

4.1.3. Experiment 2: Design.

Experiments 2a, 2b, and 2c build on the results of experiment 1 by focusing on individual feedback and probing ways to help individuals use negative feedback as effectively as positive feedback. Each experiment was conducted at a different time, but as will be described, we pool the data to analyze the results in one regression, with experiment fixed effects.

Experiment 2a uses a 2×2 between-subjects design manipulating whether feedback is provided during the surprise break midtask and whether participants receive (or, in the no feedback conditions, would have received) negative or positive feedback. In the control condition with no feedback, participants are only told how often they chose each button in the first two quarters of the game (trials 1–35 and 36–70). In the treatment conditions with feedback shown, participants are additionally shown how many points they would have earned in each quarter had there been no noise in the performance outcomes (Figure A.12 in the Online Appendix). This refined score that backs out the effects of noise is analogous to the adjustments physician report cards make for uncontrollable factors like patient or hospital factors.

Feedback valence hinges on whether the refined score accounts for good or bad luck (full details in Online Appendix Section B.2). Luck is defined by the random noise terms in the first half of the game, which are independent of participant choices. Good luck is when the noise terms boost the first-half score, whereas bad luck is when they reduce the score. Participants receiving negative feedback learn that their actual performance was worse than they perceived (being inflated by good luck) while those receiving positive feedback learn that their performance was better than they thought (being deflated by bad luck). Critically, this measure can be constructed in both the control and treatment condition because everyone experiences good or bad luck during the task. Yet only those in the treatment conditions receive the positive or negative feedback signal indicating they were lucky or unlucky.

Experiments 2b and 2c modify the design of experiment 2a by excluding the control condition with no feedback. Instead, they focus on the negative and positive feedback treatments plus additional treatments aimed at improving the use of negative feedback. Experiment 2b additionally manipulates whether feedback is paired with a hint to aid in its interpretation. Experiment 2c additionally manipulates whether participants receive a bonus based on their improvement from before to after the break (Figures A.13, (c) and (d), in the Online Appendix), on top of their pay based on overall performance. Thus, 2b examines whether issues with negative feedback arise from cognitive challenges in processing information, whereas 2c examines whether challenges are due to motivational deficits. Importantly, 2a, 2b, and 2c all include the treatment conditions in which participants are shown negative or positive feedback (with no other intervention).[24]

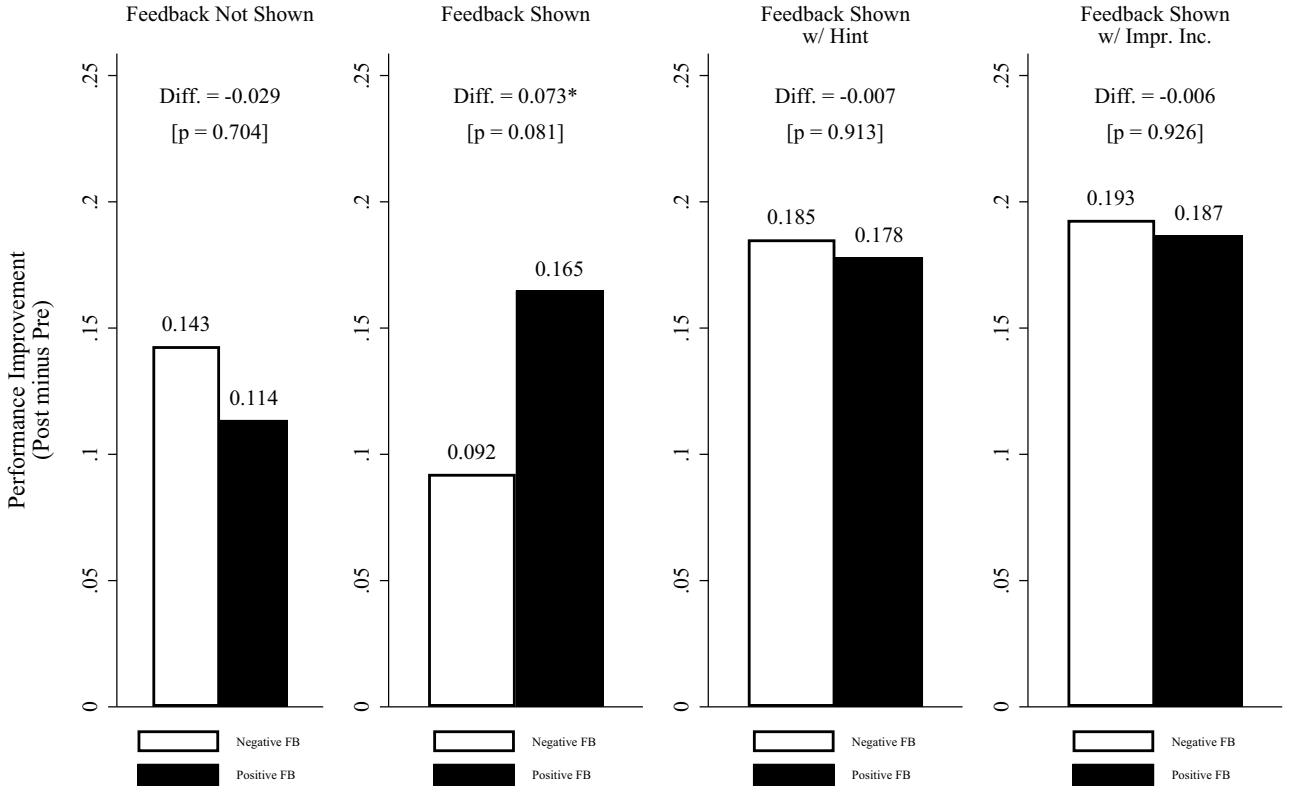
4.1.4. Experiment 2: Results.

Aggregating experiments 2a, 2b, and 2c, we analyze differences between negative and positive feedback across the four feedback interventions: (i) no feedback (experiment 2a), (ii) feedback shown (experiments 2a, 2b, and 2c), (iii) feedback with a hint (experiment 2b), and (iv) feedback with extra improvement incentives (experiment 2c). Table A.10 in the Online Appendix presents descriptive statistics. To examine the performance improvement post feedback break across conditions, we estimate the following regression with standard errors clustered by participants:

$$Performance_{ij} = Post_i \cdot PositiveFB_j \cdot FeedbackCondition_j + \varepsilon_{ij}, \quad (4)$$

where $\cdot$ indicates a fully saturated interaction and

Figure 5. Experiment 2: Performance Improvement by Feedback Valence for Each Intervention



Notes. This bar chart shows the average performance improvement from the pre to the post period in each of the eight experimental conditions in experiment 2. The values are calculated from Equation (4), and the experimental conditions are described in Section 4.1.3.

- $Performance_{ij}$ is participant j's performance on trial i, ranging from zero (choosing the worst button) to two (choosing the best button)
- $Post_i$ is one for trials 71–140 (post feedback break), and zero for trials 1–70
- $PositiveFB_j$ is one for participants who do (or would have) receive positive feedback during the break and zero otherwise
- $FeedbackCondition_j$ is zero for the control condition (no feedback shown), one for feedback without a hint or improvement incentives, two for feedback with a hint, and three for feedback with improvement incentives

Figure 5 shows the results by giving the marginal effect of $Post_i$ for each of the eight conditions (4 interventions × 2 feedback types = 8 scenarios).[25] In all eight conditions, participants improve significantly from the first to second half. More importantly, when feedback is shown without hints or improvement incentives, participants improve significantly more with positive feedback than with negative feedback. No such differences occur in the control condition in which feedback is not shown. Thus, effects are not due to the experience of good or bad luck in the first half, they are due to receiving the negative or positive feedback signal. Comparing the control condition to the treatment condition (with no hint or improvement incentives) shows that negative feedback (insignificantly) decreases performance ($p = 0.378$) while positive feedback (insignificantly) increases performance ($p = 0.437$).

The hint and improvement incentives both ameliorate the impact of feedback valence. In these conditions, individuals receiving negative and positive feedback improve at similar rates. Notably, these interventions primarily affect the use of negative feedback: In the three treatment conditions in which feedback is shown, there are no notable differences for participants who receive positive feedback. However, the hint and improvement incentives both significantly increase the performance improvement for those receiving negative feedback. More formally, within the negative feedback conditions, the effect of $Post_i$ is significantly different when $FeedbackCondition_j$ equals one versus two ($p = 0.099$) and when $FeedbackCondition_j$ equals one versus three ($p = 0.049$). The fact that the hint and improvement incentives increase performance with negative but not positive feedback suggests that they overcome issues people otherwise have in using negative feedback.

Overall, experiment 2 corroborates our findings with individual feedback. The findings also suggest interventions that practitioners like healthcare regulators

can use to encourage individuals to use negative feedback effectively. The treatment with hints indicates that feedback interventions that provide outcome data without qualitative feedback are less impactful than those combining data with explanatory insights. The treatment involving additional improvement incentives highlights the potential benefits of tying rewards to performance gains (Hermes et al. 2021). Although all participants had a performance-based incentives, extra pay tied directly to improvements seems to mitigate the adverse effects of negative feedback. However, caution is advised in interpreting these results because we only examined the hint and improvement incentives with individual performance feedback.

5. Conclusion

In this study, we find that giving physicians positive outcome feedback results in greater performance improvements than giving them negative outcome feedback. In particular, our analyses using clinical data on more than 240,000 surgeries and quasi-experimental methods show that the surprise release of a report card giving physicians information on their past surgical outcomes leads to an improvement in subsequent patient outcomes (rates of mortality and 30-day readmission) for surgeons who receive positive feedback but either no improvement or a deterioration in performance for those who receive negative feedback. These results are largely replicated in controlled experiments with laypersons and are unlikely to be explained by mean reversion, financial incentives, or changes in physicians' patient composition. Instead, the reductions in performance are more consistent with behavioral models suggesting that additional performance information can sometimes be ignored or induce suboptimal reactions like dejection, anger, confusion, stress, or a switch to ineffective task strategies.

Our findings add important nuance to prior research on the effects of physician report cards. Although these report cards have been shown to induce modest performance improvements on average (Hofer et al. 1999, Fung et al. 2008, Prang et al. 2021), our results show the average effects can mask substantial heterogeneity for disclosures at the individual level, with the report card even leading to systematically worse patient outcomes (i.e., death and 30-day readmission) in a nontrivial portion of situations. Understanding these heterogeneities provides a more complete picture of report cards and gives insights into how they can be improved. According to our results, policymakers should focus on interventions that help individuals use negative feedback more effectively. Results from our experiments suggest two such interventions: pairing the outcome feedback with qualitative information on how to use it and providing incentives directly tied to improvement.

More generally, regulators focused on improving feedback interventions in healthcare should consider behavioral factors in addition to measurement issues. Prior criticisms of report cards have largely focused on measurement issues related to a perception that they fail to perfectly risk adjust for patient and hospital factors (Jauhar 2015). The results of our study suggest that, although this concern may be valid, it may also be overstated. Even if report cards risk adjust perfectly (as in our second set of experiments), their efficacy can still be limited by behavioral responses that reduce the extent to which physicians use negative feedback well. Thus, our findings suggest that rather than focusing exclusively on better measurement, policy makers should also focus on how to accentuate beneficial behavioral responses and how to mitigate detrimental behavioral responses, especially given that the report card led to serious, unintended harm for a large portion of surgeons receiving negative feedback.

Our findings also contribute to discourse on the role of positive and negative feedback. Previous studies on relative performance information have found potential detrimental effects of both positive relative feedback (due to complacency) and negative relative feedback (due to giving up), which is inconsistent with our findings (Casas-Arce and Martinez-Jerez 2009, Berger et al. 2013). Three key aspects of our setting can explain this difference. First, prior studies that find complacency or giving up have clear nonlinear tournament incentives that encourage such behaviors. In our setting, it is unlikely that physician are competing for patients under similar tournament-like incentives (see discussion of financial incentives and demand effects in Section 3.3). In fact, the substantial downside risk from lawsuits strongly disincentivizes complacency in medicine. Second, research in settings where performance depends on finding and using effective task strategies, similar to our setting, fails to find complacency effects and instead shows monotonic effects of increasingly positive feedback on performance (Hannan et al. 2008). Third, physicians themselves may also be unique in terms of their social-comparison tendencies (Liao et al. 2016, Meeker et al. 2016). Given the competitiveness of the medical field and the type of individuals who select into that field, physicians are likely more competitive than the average person and, thus, may be less prone to complacency and giving up (see discussion of the results on competitiveness from the experiments in Section 4.1.2).

More generally, our study contributes to an emerging body of research showing that positive feedback is more effective at promoting desired behaviors than negative feedback (Lefebvre et al. 2017, Eskreis-Winkler and Fishbach 2019, Chambon et al. 2020, Gnepp et al. 2020, Erickson et al. 2022). We build on these studies by showing that even surgeons operating in highly consequential

settings respond more effectively to positive feedback, suggesting that issues with negative feedback may be even more pervasive than previously thought. Overall, accumulating evidence is beginning to suggest that, in many settings, positive reinforcers such as bonuses, kindness, and positive feedback may be more effective tools for promoting desired behaviors than negative reinforcers such as penalties, unkindness, and negative feedback (Deci et al. 1999, Eskreis-Winkler and Fishbach 2019, Gonzalez et al. 2020, Burke et al. 2023, Samet et al. 2023).

Acknowledgments

The authors thank Marika Cabral, Leemore Dafny, Qing Gong, Conor Ryan, Jennifer Kao, Yiquen, Jochen Gunter, Engy Zidane, Yiqun Chen, Christopher Whaley, Ian McCarthy, Karl Schumacher, Razvan Ghita, Jordan Samet, four anonymous conference reviewers, and workshop participants at the University of Massachusetts Amherst, University of Hawai'i at Manoa, Johannes Kepler University Linz, Maastricht University, Southern Denmark University, Rotterdam and Erasmus Universities, Bentley University, Rutgers University, and Lehigh University, and participants at the American Society of Health Economists, International Health Economics Association, Association for Public Policy Analysis and Management, and Management Accounting Section Midyear Meeting for feedback.

Endnotes

[1] There is no interaction between pre-period physician performance and feedback valence. Instead, there are independent main effects of both variables that add together.

[2] See Eyring (2020) for a notable exception. Incentives may have been particularly strong in that study because the physician grades were based on patient ratings, which may influence patient choice of provider more than ratings of surgical outcomes and because the ratings were clearly communicated to patients.

[3] Nevertheless, we explore the possibility that the report card we examine changed financial incentives differently for physicians who receive positive versus negative feedback in a way that can explain our pattern of results.

[4] Some of the studies we cite do not strictly differentiate valence from personal characteristics like ability (Hannan et al. 2008, Awaysheh et al. 2023). However, we still view them as relevant for our theoretical development because they compare people who receive good and bad performance information. Other studies focus on binary performance outcomes (success versus failure) but isolate valence effects using framing (Eskreis-Winkler and Fishbach 2019) or within-subject modeling (Lefebvre et al. 2017, Chambon et al. 2020).

[5] The surgeries are laparoscopic cholecystectomy, radical prostatectomy, transurethral prostatectomy (TURP), cervical fusion (anterior column, anterior technique), lumbar/lumbosacral fusion (posterior column, posterior technique), lumbar/lumbosacral fusion (anterior column, posterior technique), total hip replacement, and total knee replacement.

[6] Even if physicians knew the report card would not be updated, some might still alter their behavior due to professional pride and competition with peers. It would be insightful to study physician behavior over longer periods after realizing the report card would not be updated, but our data are limited in this regard.

[7] Florida provides a clean setting to examine our hypotheses because the ProPublica report card was and remains to be the only such report card for physicians in Florida.

[8] The report card scale for each physician-procedure extends over 100 units, with 0 indicating the start of the green zone and 100 the end of the red zone. Our analysis uses distance to the cutoff because the report card designers do not provide the report complication rate corresponding to any cutoff for any procedure, and this information cannot be derived. For physicians, the essential aspect is their relative position on this scale: a physician 1 unit into the yellow zone past the green-yellow cutoff is perceived as having a more favorable score than another who is 1.2 units past the same cutoff, irrespective of the actual meanings of these distances in terms of report card complication rates.

[9] To prevent overwhelming readers with excessive figures and tables, we limit discussion of the yellow-red cutoff to the nonparametric analysis with our two main models. we run the exact same tests on the yellow-red cutoff as we do for the green-yellow cutoff, but we do not report them for sake of brevity. this decision is based on the consistent finding that the yellow-red cutoff does not significantly impact physician performance.

[10] In his paper, Kolstad (2013, p. 2887) gives the following reason for physicians using observed outcomes to form beliefs over their own outcomes: "Surgeons do not … know with great certainty whether a patient is likely to have died given their underlying severity and the latest techniques and technologies … It is plausible physicians may have an indication of a given patient's severity. However, the evidence on physician difficulty in assessing probabilities, the dynamics of new technologies and techniques, and physician reliance on rules of thumb in treatment choices suggests that the objective measure of risk adjustment would provide new information relative to their existing assessments of patient severity (Frank and Zeckhauser 2007). Dziuban et al. (1994) present a case study of precisely this mechanism in response to New York State's quality report cards."

[11] We are not assuming that every physician who is better than the majority of their peers views this as positive feedback (maybe they expected to be even better than what the report card revealed). Rather, we only assume that physicians who learn they are better than the majority of their peers are more likely to view this as positive feedback than those who learn they are worse than the majority of their peers.

[12] We use all observations to avoid selectively dropping data, but in untabulated analyses, the results hold if we only analyze physician-surgeries with scores in the yellow zone.

[13] A fully saturated interaction between two regressors A and B implies the following relationship: $A \times B \equiv A + B + A \cdot B$.

[14] Including report card covariates is more relevant for the relative feedback measure of valence (to compare the effects of distribution of peers while keeping the information provided by the report card about the self-constant). However, we include it when using the individual feedback measure only for the sake of equivalence between the two analyses (though omitting it does not statistically or inferentially affect our estimates).

[15] Although the parallel trends assumption is satisfied, Panel (a) of Figures 2 and 3 does show differences in average preperiod raw outcomes. In Section 3.2.2, we conduct robustness checks to show that these differences are not driving our results. The relative performance measure depends solely on the adjusted outcomes in the report card (not raw outcomes), and there are no differences in preperiod performance when we adjust outcomes ourselves using Equation (1) (Figure 3(b)). With the individual measure, the differences in raw outcomes (Figure 2(a)) are not surprising due to the variable construction. For example, those with a high rate of raw good outcomes in the preperiod are, on average, more likely to have been benefiting from favorable performance conditions like healthy patient populations and good luck and thus more likely to have a downward adjustment (i.e.,

negative feedback) from the report card (which accounts for such factors). Thus, as expected, the preperiod differences with the individual measure decrease when we adjust patient outcomes using Equation (1) (Figure 2(b)). Yet they still persist, likely because our adjustments are not as complete as the report card adjustments because we do not have access to the full population of data that the report card does (we only have Florida).

[16] With either feedback valence measure, the event study graphs show a gradual consistent improvement in physician performance after quarter 15. This slow rise is likely due to (i) the nature of the outcomes, like in-hospital mortality and 30-day readmission, which require time for impactful changes through careful practice review and learning, thus delaying report card effects. (ii) The report card's nature as a "soft-touch" intervention, released by a nongovernmental body without formal incentives, meaning awareness likely spread gradually. Consequently, actual learning of report card scores may have happened after quarter 15 for many physicians. The uncertainty in awareness timing and report card penetration, coupled with the challenge of influencing the targeted outcomes, makes the observed effects aligning with our hypotheses indicative of strong results.

[17] That is, a physician with a raw complication rate of 4% may have a prior over their own performance of 3% after adjusting for a particularly old patient sample. This test allows us to assume that physicians are risk adjusting to some extent.

[18] That is, if a physician's observed complication rate is 4%, we randomly add or subtract up to 50% to it, thereby randomly setting their observed performance anywhere from 2% to 6%. This allows us to assume that physicians form priors over their own performance by observing the outcomes of their surgeries, but with some nontrivial error.

[19] That is, instead of only a saturated interaction between $Post_{iq} \times$ valence (D_{js}), we add a saturated interaction between $Post_{iq} \times Perf^{c}_{js}$.

[20] Specifically, the coefficients on $Post_{iq} \times$ valence are significant in all cases: low performers with individual feedback, $p < 0.001$; high performers with individual feedback, $p < 0.001$; low performers with relative feedback, $p = 0.025$; high performers with relative feedback, $p = 0.002$).

[21] For data quality, we used a CAPTCHA on page 1 along with the "Prevent multiple submissions" feature in Qualtrics. On Mechanical Turk, we used CloudResearch's "Block Low Quality Participants" (Eyal et al. 2021) and only recruited people who had completed over 100 HITs with a minimum 95% approval rate.

[22] In experiment 1, random choices would yield 3,500 points on average, but participants averaged 3,570.81 points ($p < 0.001$). In experiments 2a, 2b, and 2c, random choices would yield 4,900 points, but participants averaged 5,031.75, 5,002.95, and 5,003.86 points, respectively (all $p < 0.001$).

[23] Table A.8 shows full details on the marginal effects of Post and Table A.9 shows the standard regression output.

[24] Untabulated analyses show that the effects of negative versus positive feedback in the common treatment condition did not vary across studies (smallest p value is 0.805).

[25] Table A.12 gives standard regression coefficients and Table A.11 gives complete statistics for marginal effects.

References

Abraham J, Feldman R, Carlin C (2004) Understanding employee awareness of healthcare quality information: How can employers benefit? *Health Services Res.* 39(6p1):1799–1816.

AHRQ (2017) About the national quality strategy. Accessed October 15, 2022, https://www.ahrq.gov/workingforquality/about/index.html.

Anderson G, Hussey PS (2001) Comparing health system performance in OECD countries. *Health Affairs* 20(3):219–232.

Anderson SW, Kimball A (2019) Evidence for the feedback role of performance measurement systems. *Management Sci.* 65(9):4385–4406.

Andrabi T, Das J, Khwaja AI (2017) Report cards: The impact of providing school and child test scores on educational markets. *Amer. Econom. Rev.* 107(6):1535–1563.

Audia PG, Locke EA (2003) Benefiting from negative feedback. *Human Resource Management Rev.* 13(4):631–646.

Awaysheh A, Bonet R, Ortega J (2023) Performance feedback and productivity: Evidence from a field experiment. *Production Oper. Management* 32(1):98–115.

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

Bailey R (2014) The problem with praise. Accessed October 15, 2022, https://www.psychologytoday.com/intl/blog/smart-moves/201411/the-problem-praise.

Balcazar F, Hopkins BL, Suarez Y (1985) A critical, objective review of performance feedback. *J. Organ. Behav. Management* 7(3–4):65–89.

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. *Rev. General Psych.* 5(4):323–370.

Berger L, Klassen KJ, Libby T, Webb A (2013) Complacency and giving up across repeated tournaments: Evidence from the field. *J. Management Accounting Res.* 25(1):143–167.

Berner ES, Graber ML (2008) Overconfidence as a cause of diagnostic error in medicine. *Amer. J. Medicine* 121(5):S2–S23.

Brockner J, Derr WR, Laing WN (1987) Self-esteem and reactions to negative feedback: Toward greater generalizability. *J. Res. Personality* 21(3):318–333.

Brown JD (2010) High self-esteem buffers negative feedback: Once more with feeling. *Cognition Emotion* 24(8):1389–1404.

Bryan CJ, Tipton E, Yeager DS (2021) Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behav.* 5(8):980–989.

Burke J, Towry KL, Young D, Zureich J (2023) Ambiguous sticks and carrots: The effect of contract framing and payoff ambiguity on employee effort. *Accounting Rev.* 98(1):139–162.

Callaway B, Sant'Anna PHC (2021) Difference-in-differences with multiple time periods. *J. Econometrics* 225(2):200–230.

Calonico S, Cattaneo MD, Farrell MH, Titiunik R (2017) rdrobust: Software for regression-discontinuity designs. *Stata J.* 17(2):372–404.

Cannon MD, Edmondson AC (2005) Failing to learn and learning to fail (intelligently): How great organizations put failure to work to innovate and improve. *Long Range Planning* 38(3):299–319.

Carver CS, Scheier MF (2001) *On the Self-Regulation of Behavior* (Cambridge University Press, Cambridge, UK).

Casas-Arce P, Martinez-Jerez FA (2009) Relative performance compensation, contests, and dynamic incentives. *Management Sci.* 55(8):1306–1320.

Casas-Arce P, Lourenco SM, Martinez-Jerez F (2017) The performance effect of feedback frequency and detail: Evidence from a field experiment in customer satisfaction. *J. Accounting Res.* 55(5):1051–1088.

Chab'e-Ferret S (2015) Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *J. Econometrics* 185(1):110–123.

Chambon V, Thero H, Vidal M, Vandendriessche H, Haggard P, Palminteri S (2020) Information about action outcomes differentially affects learning from self-determined vs. imposed choices. *Nature Human Behav.* 4(10):1067–1079.

Chen Y-C, Hung M, Wang Y (2018) The effect of mandatory CSR disclosure on firm profitability and social externalities: Evidence from China. *J. Accounting Econom.* 65(1):169–190.

Chou SY, Deily ME, Li S, Lu Y (2014) Competition and the impact of online hospital report cards. *J. Health Econom.* 34:42–58.

Christensen HB, Floyd E, Liu LY, Maffett M (2017) The real effects of mandated information on social responsibility in financial

reports: Evidence from mine-safety records. *J. Accounting Econom.* 64(2–3):284–304.

Christensen-Szalanski JJ, Bushyhead JB (1981) Physicians' use of probabilistic information in a real clinical setting. *J. Experiment. Psych. Human Perception Performance* 7(4):928.

Christian B, Griffiths T (2016) *Algorithms to Live By: The Computer Science of Human Decisions* (Macmillan, New York).

Christianson JB, Volmar KM, Alexander J, Scanlon DP (2010) A report card on provider report cards: Current status of the healthcare transparency movement. *J. General Internal Medicine* 25(11): 1235–1241.

Christianson J, Maeng D, Abraham J, Scanlon DP, Alexander J, Mittler J, Finch M (2014) What influences the awareness of physician quality information? Implications for Medicare. *Medicare Medicaid Res. Rev.* 4(2):E1–E15.

Clemens J, Gottlieb JD (2014) Do physicians' financial incentives affect medical treatment and patient health? *Amer. Econom. Rev.* 104(4): 1320–1349.

Clement JP, Bazzoli GJ, Zhao M (2012) Nursing home price and quality responses to publicly reported quality information. *Health Services Res.* 47(1pt1):86–105.

Compte O, Postlewaite A (2004) Confidence-enhanced performance. *Amer. Econom. Rev.* 94(5):1536–1557.

Coutts A (2019) Good news and bad news are still news: Experimental evidence on belief updating. *Experiment. Econom.* 22(2):369–395.

Croskerry P (2003) The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad. Medicine* 78(8):775–780.

Cunningham S (2021) *Causal Inference: The Mixtape* (Yale University Press, New Haven, CT).

Deci EL, Koestner R, Ryan RM (1999) A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psych. Bull.* 125(6):627.

Dorfman HM, Bhui R, Hughes BL, Gershman SJ (2019) Causal inference about good and bad outcomes. *Psych. Sci.* 30(4):516–525.

Dunt D, Prang K-H, Sabanovic H, Kelaher M (2018) The impact of public performance reporting on market share, mortality, and patient mix outcomes associated with coronary artery bypass grafts and percutaneous coronary interventions (2000–2016): A systematic review and meta-analysis. *Medical Care* 56(11):956.

Dziuban SW Jr, McIlduff JB, Miller SJ, Dal Col RH (1994) How a New York cardiac surgery program uses outcomes data. *Annals Thoracic Surgery* 58(6):1871–1876.

Eddy DM (1982) Probabilistic reasoning in clinical medicine: Problems and opportunities. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, UK), 249–267.

Edmondson AC (2011) Strategies for learning from failure. *Harvard Bus. Rev.* 89(4):48–55.

Eil D, Rao JM (2011) The good news-bad news effect: Asymmetric processing of objective information about yourself. *Amer. Econom. J.* 3(2):114–138.

Elder J, Davis T, Hughes BL (2022) Learning about the self: Motives for coherence and positivity constrain learning from self-relevant social feedback. *Psych. Sci.* 33(4):629–647.

Elixhauser A, Steiner C, Harris DR, Coffey RM (1998) Comorbidity measures for use with administrative data. *Medical Care* 36(1): 8–27.

Elstein AS, Shulman LS, Sprafka SA (1981) Medical problem-solving. *Acad. Medicine* 56(1):75–76.

Engelberg S, Pierce O (2015) Our rebuttal to rand's critique of surgeon scorecard. Accessed October 15, 2022, https://www.propublica.org/article/our-rebuttal-to-rands-critique-of-surgeon-scorecard.

Epstein AJ (2010) Effects of report cards on referral patterns to cardiac surgeons. *J. Health Econom.* 29(5):718–731.

Erickson D, Holderness DK Jr, Olsen KJ, Thornock TA (2022) Feedback with feeling? How emotional language in feedback affects individual performance. *Accounting Organ. Soc.* 99:101329.

Eskreis-Winkler L, Fishbach A (2019) Not learning from failure—The greatest failure of all. *Psych Sci.* 30(12):1733–1744.

Evans MF (2016) The clean air act watch list: An enforcement and compliance natural experiment. *J. Assoc. Environment. Resources Econom.* 3(3):627–665.

Eyal P, David R, Andrew G, Zak E, Ekaterina D (2021) Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods*, 1–20.

Eyring H (2020) Disclosing physician ratings: Performance effects and the difficulty of altering ratings consensus. *J. Accounting Res.* 58(4):1023–1067.

Eyring H, Ferguson PJ, Koppers S (2021) Less information, more comparison, and better performance: Evidence from a field experiment. *J. Accounting Res.* 59(2):657–711.

Festinger L (1954) A theory of social comparison processes. *Human Relations* 7(2):117–140.

Finkelstein SR, Fishbach A (2012) Tell me what I did wrong: Experts seek and respond to negative feedback. *J. Consumer Res.* 39(1): 22–38.

Fishbach A, Eyal T, Finkelstein SR (2010) How positive and negative feedback motivate goal pursuit. *Soc. Personality Psych. Compass* 4(8):517–530.

Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL (2003) The implications of regional variations in Medicare spending. Part 1: The content, quality, and accessibility of care. *Annals Internal Medicine* 138(4):273–287.

Fong CJ, Patall EA, Vasquez AC, Stautberg S (2019) A meta-analysis of negative feedback on intrinsic motivation. *Ed. Psych. Rev.* 31: 121–162.

Frank RG, Zeckhauser RJ (2007) Custom-made versus ready-to-wear treatments: Behavioral propensities in physicians' choices. *J. Health Econom.* 26(6):1101–1127.

Frederickson JR (1992) Relative performance information: The effects of common uncertainty and contract type on agent effort. *Accounting Rev.* 67(4):647–669.

Friedberg MW, Bilimoria KY, Pronovost PJ, Shahian DM, Damberg CL, Zaslavsky AM (2016) Response to Propublica's rebuttal of our critique of the surgeon scorecard. *Rand Health Quart.* 6(1):4.

Friesen J, Javdani M, Smith J, Woodcock S (2012) How do school 'report cards' affect school choice decisions? *Canadian J. Econom.* 45(2):784–807.

Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG (2008) Systematic review: The evidence that publishing patient care performance data improves quality of care. *Annals Internal Medicine* 148(2):111–123.

Gallani S, Kajiwara T, Krishnan R (2020) Value of new performance information in healthcare: Evidence from Japan. *Internat. J. Health Econom. Management* 20(4):319–357.

Garcia SM, Tor A, Schiff TM (2013) The psychology of competition: A social comparison perspective. *Perspective Psych. Sci.* 8(6):634–650.

Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S (2007) Helping doctors and patients make sense of health statistics. *Psych. Sci. Public Interest* 8(2):53–96.

Gilovich T (1983) Biased evaluation and persistence in gambling. *J. Personality Soc. Psych.* 44(6):1110.

Gnepp J, Klayman J, Williamson IO, Barlas S (2020) The future of feedback: Motivating performance improvement through future-focused feedback. *PLoS One* 15(6):e0234444.

Goldsmith J, Kaufman N, Burns L (2016) The tangled hospital-physician relationship. Accessed June 26, 2022, https://www.healthaffairs.org/content/forefront/tangled-hospital-physician-relationship.

Gonzalez GC, Hoffman VB, Moser DV (2020) Do effort differences between bonus and penalty contracts persist in labor markets? *Accounting Rev.* 95(3):205–222.

Goodman-Bacon A (2021) Difference-in-differences with variation in treatment timing. *J. Econometrics* 225(2):254–277.

Grabowski DC, Town RJ (2011) Does information matter? Competition, quality, and the impact of nursing home report cards. *Health Services Res.* 46(6pt1):1698–1719.

Grossman Z, Owens D (2012) An unlucky feeling: Overconfidence and noisy feedback. *J. Econom. Behav. Organ.* 84(2):510–524.

Hackbarth G, Reischauer R, Mutti A (2008) Collective accountability for medical care—Toward bundled Medicare payments. *New England J. Medicine* 359(1):3–5.

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

Harackiewicz JM (1979) The effects of reward contingency and performance feedback on intrinsic motivation. *J. Personality Soc. Psych.* 37(8):1352.

Hattie J, Timperley H (2007) The power of feedback. *Rev. Ed. Res.* 77(1):81–112.

Hermes H, Huschens M, Rothlauf F, Schunk D (2021) Motivating low-achievers—Relative performance feedback in primary schools. *J. Econom. Behav. Organ.* 187:45–59.

Hirth RA, Banaszak-Holl JC, Fries BE, Turenne MN (2003) Does quality influence consumer choice of nursing homes? Evidence from nursing home to nursing home transfers. *INQUIRY* 40(4):343–361.

Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG (1999) The unreliability of individual physician report cards for assessing the costs and quality of care of a chronic disease. *JAMA* 281(22):2098–2105.

Hogarth RM, Gibbs BJ, McKenzie CR, Marquis MA (1991) Learning from feedback: Exactingness and incentives. *J. Experiment. Psych. Learn. Memory Cognition* 17(4):734.

Ilies R, De Pater IE, Judge T (2007) Differential affective reactions to negative and positive feedback, and the role of self-esteem. *J. Management Psych.* 22(6):590–609.

Imbens GW, Lemieux T (2008) Regression discontinuity designs: A guide to practice. *J. Econometrics* 142(2):615–635.

Jacobson M, Earle CC, Price M, Newhouse JP (2010) How Medicare's payment cuts for cancer chemotherapy drugs changed patterns of treatment. *Health Affairs* 29(7):1391–1399.

Jaffe TA, Hasday SJ, Dimick JB (2016) Power outage—Inadequate surgeon performance measures leave patients in the dark. *JAMA Surgery* 151(7):599–600.

Jauhar S (2015) Opinion—Giving doctors grades. Accessed October 15, 2022, https://www.nytimes.com/2015/07/22/opinion/giving-doctors-grades.html.

Jencks SF, Williams MV, Coleman EA (2009) Rehospitalizations among patients in the Medicare fee-for-service program. *New England J. Medicine* 360(14):1418–1428.

Jin GZ, Leslie P (2003) The effect of information on product quality: Evidence from restaurant hygiene grade cards. *Quart. J. Econom.* 118(2):409–451.

Jin GZ, Leslie P (2009) Reputational incentives for restaurant hygiene. *Amer. Econom. J.* 1(1):237–267.

Jin GZ, Sorensen AT (2006) Information and consumer choice: The value of publicized health plan ratings. *J. Health Econom.* 25(2):248–275.

Jung K, Feldman R, Scanlon D (2011) Where would you go for your next hospitalization? *J. Health Econom.* 30(4):832–841.

Kahneman D, Tversky A (2013) Prospect theory: An analysis of decision under risk. *Handbook of the Fundamentals of Financial Decision Making: Part I* (World Scientific, Singapore), 99–127.

Kaplan RS, Norton DP (1996) Using the balanced scorecard as a strategic management system. *Harvard Bus. Rev.* (January–February): 37–48.

Kc D, Staats BR, Gino F (2013) Learning from my success and from others' failure: Evidence from minimally invasive cardiac surgery. *Management Sci.* 59(11):2435–2449.

Khullar D, Dave A Chokshi RK, Reddy A, Basu K, Conway PH, Rajkumar R (2015) Behavioral economics and physician compensation—Promise and challenges. *New England J. Medicine* 372(24):2281–2283.

Klinger E, Barta SG, Maxeiner ME (1980) Motivational correlates of thought content frequency and commitment. *J. Personality Soc. Psych.* 39(6):1222.

Kluger AN, DeNisi A (1996) The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psych Bull.* 119(2):254.

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

Lefebvre G, Lebreton M, Meyniel F, Bourgeois-Gironde S, Palminteri S (2017) Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behav.* 1(4):1–9.

Leuz C, Wysocki PD (2016) The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. *J. Accounting Res.* 54(2):525–622.

Levitt SD, List JA, Syverson C (2013) Toward an understanding of learning by doing: Evidence from an automobile assembly plant. *J. Political Econom.* 121(4):643–681.

Levy PE, Williams JR (2004) The social context of performance appraisal: A review and framework for the future. *J. Management* 30(6):881–905.

Liao JM, Fleisher LA, Navathe AS (2016) Increasing the value of social comparisons of physician performance using norms. *JAMA* 316(11):1151–1152.

Locke EA, Latham GP (1990) *A Theory of Goal Setting & Task Performance* (Prentice-Hall, Upper Saddle River, NJ).

Loewenstein G, Molnar A (2018) The renaissance of belief-based utility in economics. *Nature Human Behav.* 2(3):166–167.

Loftus S, Tanlu LJ (2018) Because of "because": Examining the use of causal language in relative performance feedback. *Accounting Rev.* 93(2):277–297.

Lourenço SM (2016) Monetary incentives, feedback, and recognition—Complements or substitutes? Evidence from a field experiment in a retail services company. *Accounting Rev.* 91(1):279–297.

MacLean CH, Kerr EA, Qaseem A (2018) Time out—Charting a path for improving performance measurement. *New England J. Medicine* 378(19):1757–1761.

Manthei K, Sliwka D, Vogelsang T (2022) Information, incentives, and attention: A field experiment on the interaction of management controls. *Accounting Rev.* 98(5):455–479.

McNeil BJ, Pauker SG, Sox HC, Tversky A (1982) On the elicitation of preferences for alternative therapies. *New England J. Medicine* 306(21):1259–1262.

Meeker D, Linder JA, Fox CR, Friedberg MW, Persell SD, Goldstein NJ, Knight TK, et al. (2016) Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices: A randomized clinical trial. *JAMA* 315(6):562–570.

Miller S, Johnson N, Wherry LR (2021) Medicaid and mortality: New evidence from linked survey and administrative data. *Quart. J. Econom.* 136(3):1783–1829.

Moon K (2021) The real reason why it's harder than ever to get into medical school—And what aspiring physicians can do to improve their chances. Accessed October 15, 2022, https://www.forbes.com/sites/kristenmoon/2021/06/17/the-real-reason-why-its-harder-than-ever-to-get-into-medical-school-and-what-aspiring-physi.

Moore BJ, White S, Washington R, Coenen N, Elixhauser A (2017) Identifying increased risk of readmission and in-hospital mortality using hospital administrative data. *Medical Care* 55(7): 698–705.

Mukamel DB, Weimer DL, Spector WD, Ladd H, Zinn JS (2008) Publication of quality report cards and trends in reported quality measures in nursing homes. *Health Services Res.* 43(4):1244–1262.

Mullahy J, Norton EC (2022) Why transform y? A critical assessment of dependent-variable transformations in regression models for skewed and sometimes-zero outcomes. Technical report, National Bureau of Economic Research, Cambridge, MA.

Murphy B (2018) *Residency Match: The 7 Most Competitive Medical Specialties* (American Medical Association, Chicago).

Nease AA, Mudgett BO, Quinones MA (1999) Relationships among feedback sign, self-efficacy, and acceptance of performance feedback. *J. Appl. Psych.* 84(5):806.

Posen HE, Levinthal DA (2012) Chasing a moving target: Exploitation and exploration in dynamic environments. *Management Sci.* 58(3): 587–601.

Prang K-H, Maritz R, Sabanovic H, Dunt D, Kelaher M (2021) Mechanisms and impact of public reporting on physicians and hospitals' performance: A systematic review (2000–2020). *PLoS One* 16(2):e0247297.

Redelmeier DA, Shafir E (1995) Medical decision making in situations that offer multiple alternatives. *JAMA* 273(4):302–305.

Redelmeier DA, Tversky A (1990) Discrepancy between medical decisions for individual patients and for groups. Shafir E, ed. *A Bradford Book* (The MIT Press, Cambridge, MA).

Rosenbaum L (2015) Scoring no goal—Further adventures in transparency. *New England J. Medicine* 373(15):1385–1388.

Ruzzene M, Noller P (1986) Feedback motivation and reactions to personality interpretations that differ in favorability and accuracy. *J. Personality Soc. Psych.* 51(6):1293.

Samet J, Schuhmacher K, Towry KL, Zureich J (2023) Reciprocity over time: Do employees respond more to kind or unkind controls? Preprint, submitted March 6, https://dx.doi.org/10.2139/ssrn.4371449.

Scanlon DP, Shi Y, Bhandari N, Christianson JB (2015) Are healthcare quality 'report cards' reaching consumers? Awareness in the chronically ill population. *Amer. J. Management Care* 21(3):236–244.

Schneider EC, Epstein AM (1998) Use of public performance reports: A survey of patients undergoing cardiac surgery. *JAMA* 279(20): 1638–1642.

Scott KW, Orav EJ, Cutler DM, Jha AK (2017) Changes in hospital–physician affiliations in us hospitals and their effect on quality of care. *Ann. Internal Medicine* 166(1):1–8.

Shi Y, Scanlon DP, Bhandari N, Christianson JB (2017) Is anyone paying attention to physician report cards? The impact of increased availability on consumers' awareness and use of physician quality information. *Health Services Res.* 52(4):1570–1589.

Sinaiko AD, Eastman D, Rosenthal MB (2012) How report cards on physicians, physician groups, and hospitals can have greater impact on consumer choices. *Health Affairs* 31(3):602–611.

Slaughter MJ (2001) Trade liberalization and per capita income convergence: A difference-in-differences analysis. *J. Internat. Econom.* 55(1):203–228.

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

Sun L, Abraham S (2021) Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econometrics* 225(2):175–199.

Swift V, Peterson JB (2018) Improving the effectiveness of performance feedback by considering personality traits and task demands. *PLoS One* 13(5):e0197810.

Syed M (2015) *Black Box Thinking: Why Most People Never Learn from Their Mistakes: But Some Do* (Penguin, London).

Tafkov ID (2012) Private and public relative performance information under different compensation contracts. *Accounting Rev.* 88(1):327–350.

Tikkanen R, Abrams MK (2020) *US Healthcare from a Global Perspective, 2019: Higher Spending, Worse Outcomes* (The Commonwealth Fund, New York).

Uhrig JD, Short PF (2002) Testing the effect of quality reports on the health plan choices of medicare beneficiaries. *INQUIRY* 39(4): 355–371.

Van-Dijk D, Kluger AN (2004) Feedback sign effect on motivation: Is it moderated by regulatory focus? *Appl. Psych.* 53(1): 113–135.

VanLare JM, Blum JD, Conway PH (2012) Linking performance with payment: Implementing the physician value-based payment modifier. *JAMA* 308(20):2089–2090.

Weiner B (1985) "Spontaneous" causal thinking. *Psych. Bull.* 97(1):74.

Wennberg JE (2004) Practice variations and healthcare reform: Connecting the dots: A focus on medical error is preventing sufficient focus on improving the quality of patient decision making to reduce practice variations (and costs) in today's healthcare system. *Health Affairs* 23(Suppl 2): VAR–140–VAR–144.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

Zimmermann F (2020) The dynamics of motivated beliefs. *Amer. Econom. Rev.* 110(2):337–361.