PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

When peer comparison information harms physician well-being

Joseph S. Reiff[a,1], Justin C. Zhang[b,1], Jana Gallus[a], Hengchen Dai[a], Nathaniel M. Pedley[c], Sitaram Vangala[c], Richard K. Leuchter[c], Gregory Goshgarian[d], Craig R. Fox[a], Maria Han[c], and Daniel M. Croymans[c,2]

Edited by Timothy Wilson, University of Virginia, Charlottesville, VA; received November 30, 2021; accepted April 23, 2022

Policymakers and business leaders often use peer comparison information—showing people how their behavior compares to that of their peers—to motivate a range of behaviors. Despite their widespread use, the potential impact of peer comparison interventions on recipients' well-being is largely unknown. We conducted a 5-mo field experiment involving 199 primary care physicians and 46,631 patients to examine the impact of a peer comparison intervention on physicians' job performance, job satisfaction, and burnout. We varied whether physicians received information about their preventive care performance compared to that of other physicians in the same health system. Our analyses reveal that our implementation of peer comparison did not significantly improve physicians' preventive care performance, but it did significantly decrease job satisfaction and increase burnout, with the effect on job satisfaction persisting for at least 4 mo after the intervention had been discontinued. Quantitative and qualitative evidence on the mechanisms underlying these unanticipated negative effects suggest that the intervention inadvertently signaled a lack of support from leadership. Consistent with this account, providing leaders with training on how to support physicians mitigated the negative effects on well-being. Our research uncovers a critical potential downside of peer comparison interventions, highlights the importance of evaluating the psychological costs of behavioral interventions, and points to how a complementary intervention—leadership support training—can mitigate these costs.

peer comparison | well-being | healthcare | field experiment

Significance

Motivating physicians to adhere to medical best practices is a constant concern for health system leaders and policymakers. Meanwhile, burnout rates among physicians are rising—often resulting in mental health problems, job turnover, and higher healthcare costs. In our study, a commonly used behavioral intervention—informing physicians about how their performance compares to that of their peers—has no statistically significant impact on performance. However, it does decrease physicians' job satisfaction and increase burnout. We uncover one mechanism behind this backfiring effect, namely, that the intervention may signal a lack of leadership support. Consistent with this account, we find that training leaders to offer support offsets the negative impact. We discuss lessons for the design, implementation, and evaluation of behavioral interventions and policies.

Many behavioral change interventions leverage peer comparison information, which involves showing people how their behavior compares to that of their peers. Peer comparison interventions have successfully improved educational outcomes (1), reduced energy consumption (2), boosted voter turnout (3), increased charitable giving (4), and bolstered employee productivity (5). Within healthcare systems, peer comparison interventions targeting physicians have curbed overprescribing of antibiotics (6), improved emergency department efficiency (7), and increased adherence to best practices (8). Previous research has primarily focused on how peer comparison interventions affect targeted behaviors. Yet, by only focusing on these behaviors, researchers and policymakers risk overlooking an important, less-visible class of outcomes, namely, recipient well-being.

The original goal of the current research was to evaluate whether a newly introduced peer comparison intervention would improve physicians' preventive care performance. In a natural field experiment within a large hospital system, we found no evidence of such an effect on physician performance. However, we observed an unexpected negative impact of the peer comparison intervention on physicians' job satisfaction and burnout. The primary goal of this paper is to understand these harmful effects so that they can be avoided in the future.

Recent research suggests that peer comparison information can be aversive to recipients (9). In particular, being compared to higher ranked peers can be discouraging (10–12), resulting in feelings of shame (13) or stress (14). Extending prior work that has focused on immediate affective reactions to upward social comparisons, we theorize that, when implemented in organizational contexts, peer comparison interventions can elicit another psychological process and impose long-term psychological costs. We propose that the use of peer comparison interventions can alter workers' perceptions of and relationships with the leaders implementing the intervention as they try to make sense of how and why this information is being presented to them (15). Workers may perceive their leaders' use of the intervention as reflecting inadequate leadership support if workers deem that the intervention's design and implementation violate existing norms of cooperation (5, 16) or contradict workers' beliefs about what constitutes appropriate performance feedback. Given that leadership support is key to work-related well-being* (17–20), job satisfaction and burnout may be harmed by the use of peer comparison interventions.

*We use the phrase "work-related well-being," or "well-being" for short, to refer to employees' job satisfaction and burnout.

Author contributions: J.S.R., J.C.Z., J.G., H.D., N.M.P., R.K.L., G.G., C.R.F., M.H., and D.M.C. designed research; J.S.R., J.C.Z., J.G., H.D., N.M.P., and D.M.C. performed research; J.S.R. and S.V., analyzed data; J.S.R. and J.C.Z. wrote the paper; and J.S.R., J.C.Z., J.G., H.D., N.M.P., S.V., R.K.L., G.G., C.R.F., M.H., and D.M.C. revised the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]J.S.R. and J.C.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: dcroymans@mednet.ucla.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2121730119/-/DCSupplemental.

Published July 14, 2022.

These dynamics are particularly important to examine within the healthcare context, where public health leaders must balance dual objectives. As health insurance plans place greater weight on optimizing "healthcare quality" metrics, health systems across the United States are increasingly tracking physician behavior and implementing behavioral interventions (e.g., using peer comparison information) in an attempt to improve performance on these metrics (21, 22). Even Medicare has administered large-scale programs that use peer comparison information (23, 24). Concurrently, almost half of physicians in the United States report experiencing burnout (25), which is associated with greater turnover, reduced job performance, increased alcohol abuse, and higher rates of suicide (20, 26–29)—estimated to cost the US healthcare system $5 billion annually (30, 31).

Results

We conducted a 5-mo field experiment (from November 2019 through March 2020) in partnership with University of California, Los Angeles (UCLA) Health to examine the impact of a peer comparison intervention on both physicians' job performance and well-being. The experiment involved 199 primary care physicians (PCPs) and their 46,631 patients. PCPs were cluster randomized at the clinic level to one of three study conditions, as follows: control (condition 1), peer comparison (condition 2), or peer comparison and leadership training (condition 3). PCPs in all conditions received monthly emails from UCLA Health's department leadership with feedback about their preventive care performance. Their performance was summarized with a "health maintenance (HM) completion rate," which reflects the proportion of recommended preventive care measures, such as routine screenings, that were completed by their patients in the previous 3 mo. The emails in the control condition only contained feedback about the PCP's personal score. The emails in the peer comparison condition also contained a list of the month's "Top 25 Primary Care Physicians" as well as information about where the PCP fell in the performance distribution. PCPs in the peer comparison and leadership training condition received the same emails as those used in the peer comparison condition, but leaders at each clinic also participated in training on how to support their physicians' preventive care performance. See *Materials and Methods* and SI Appendix for more information.

Order Rate of Preventive Screening Examinations. For each patient who visited a PCP in our experiment, we tracked the share of recommended preventive measures that were ordered by their PCP within the 7 d following the visit. This is our primary preregistered outcome. The average order rates were 9.4% in the control condition (SD = 25.4%), 10.5% in the peer comparison condition (SD = 26.5%), and 9.9% in the peer comparison and leadership training condition (SD = 26.0%). Following our preregistered analysis plan, we first compared PCPs' order rates between the control condition and the conditions containing peer comparison information (conditions 2 and 3) and found no statistically significant difference ($P = 0.143$). As an exploratory analysis, we also compared the order rates between condition 1 and condition 2 but still did not find any statistically significant differences ($P = 0.324$). The regression tables for these analyses are reported in *SI Appendix*, Section 8.

Previous research suggests that the impact of peer comparison may depend on baseline performance (32–34), discouraging low performers while encouraging high performers. However, our post hoc analysis found no evidence that the estimated effect of the peer comparison intervention on order rates was moderated by PCPs' baseline performance (i.e., the HM completion rate displayed in the first intervention email). See *SI Appendix*, Section 9 for details.

Job Satisfaction and Burnout. Next, we examined differences between conditions in our two well-being outcomes, namely, job satisfaction and burnout, which were measured by UCLA Health in quarterly surveys. We first confirmed that job satisfaction and burnout were balanced across conditions in the baseline period before the experiment started (October 2019; F-test for joint significance: job satisfaction, $P = 0.432$; burnout, $P = 0.134$). We then evaluated the effects of our interventions on job satisfaction and burnout at the end of the 5-mo experimental period (April 2020). The regression-estimated treatment effects are displayed in Figs. 1 and 2. Since both the peer comparison and the leadership support training interventions could separately impact well-being, we first evaluated the effects of peer comparison alone (comparing condition 2 with condition 1). We then tested the effects of adding leadership training (condition 3 vs. condition 2). Compared to the control condition (job satisfaction, M = 5.47, SD = 0.91; burnout, M = 1.93, SD = 0.73), the peer comparison intervention (condition 2) significantly decreased job satisfaction (M = 4.95, SD = 1.48; $\beta = -0.55$, 95% CI = [−1.01, −0.09], $P = 0.021$, d = 0.42) and increased burnout (M = 2.47, SD = 0.96; $\beta = 0.33$, 95% CI = [0.03, 0.63], $P = 0.031$, d = 0.64). In contrast, PCPs who received leadership support training combined with peer comparison (condition 3) experienced significantly higher job satisfaction (M = 5.29, SD = 1.27; $\beta = 0.45$, 95% CI = [0.02, 0.88], $P = 0.044$, d = 0.25) and lower burnout (M = 2.09, SD = 0.84; $\beta = -0.44$, 95% CI = [−0.79, −0.09], $P = 0.016$, d = 0.42) than PCPs who received the peer comparison intervention alone (condition 2). The results remained statistically significant at the 5% level after a twofold Holm–Bonferroni correction that adjusted for multiple hypothesis testing due to simultaneously comparing conditions 2 vs. 1 and conditions 3 vs. 2 (conditions 2 vs. 1: job satisfaction adjusted $P = 0.042$, burnout adjusted $P = 0.032$; conditions 3 vs. 2: job satisfaction adjusted $P = 0.044$, burnout adjusted $P = 0.032$). Finally, we found no significant differences in job satisfaction or burnout between condition 3 and the control condition ($P = 0.509$ and $P = 0.364$, respectively; *SI Appendix*, Section 10).

Robustness Checks and Secondary Analyses. Our aforementioned results about physician well-being were robust to excluding controls for physician characteristics or including the number of positive COVID-19 cases each PCP encountered as a control. Additionally, in a post hoc placebo test, we confirmed no statistically significant effect of the peer comparison intervention alone or the leadership training on an outcome that we would not expect to be impacted by the interventions (perceived proficiency with the electronic health record system). Finally, we found no evidence that the negative effects of the peer comparison intervention on well-being were moderated by PCPs' baseline performance. See of *SI Appendix*, Section 11 for these robustness checks and secondary analyses.

Treatment Effect Persistence. To explore the persistence of our interventions' treatment effects, we analyzed survey responses collected 4 mo after the interventions had been discontinued (July 2020; see Figs. 1 and 2 for the regression-estimated treatment effects and *SI Appendix*, Section 12 for more details). The negative effect of the peer comparison intervention (condition 2 vs. control) on job satisfaction remained significant (control condition: M = 5.22, SD = 1.07; condition 2:

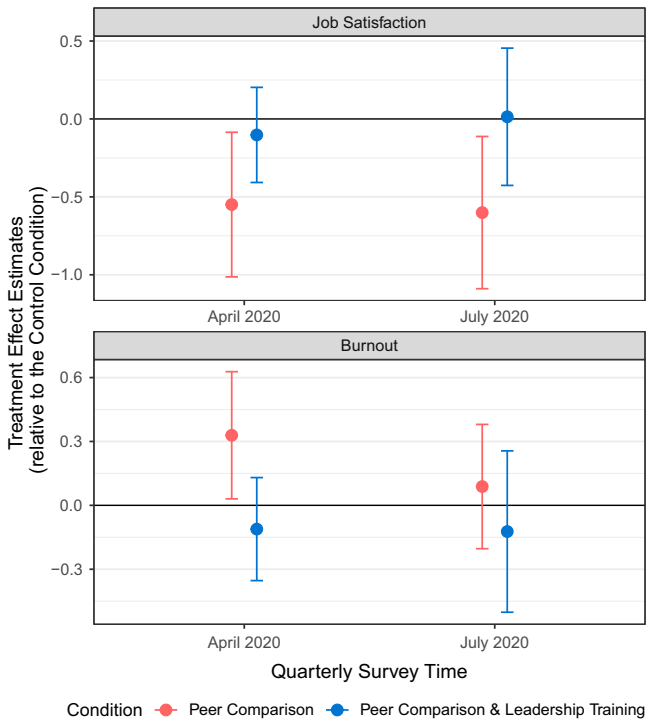

Fig. 1. Treatment effect estimates on job satisfaction and burnout. The blue and red dots reflect the estimated treatment effects of the respective conditions (vs. Control Condition) on job satisfaction (upper panel) and burnout (lower panel). Error bars reflect 95% confidence intervals.

M = 4.64, SD = 1.51; β = −0.60, 95% CI = [−1.09, −0.12], $P = 0.017$, d = 0.45). Moreover, PCPs who received leadership support training combined with peer comparison (condition 3) persistently experienced significantly higher job satisfaction (condition 3: M = 5.21, SD = 1.38; β = 0.62, 95% CI = [0.14, 1.09],

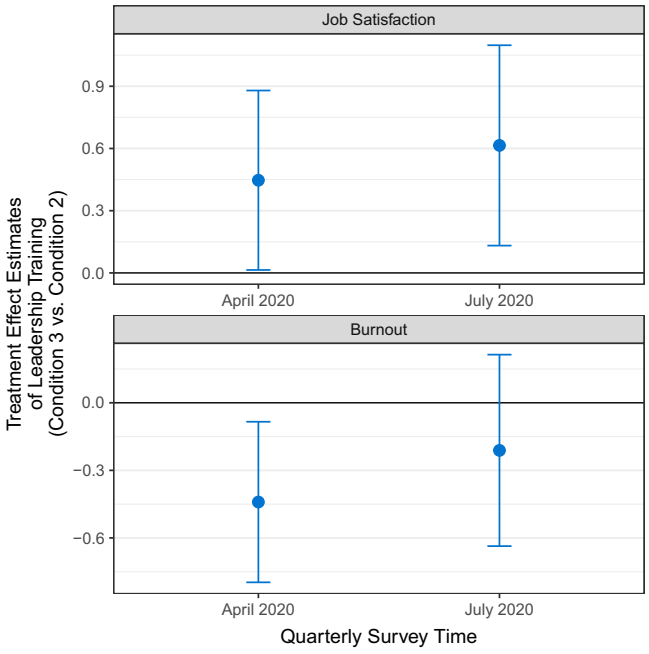

Fig. 2. Treatment effect estimates of adding leadership support training to the peer comparison intervention. The blue dots reflect the estimated treatment effects on job satisfaction (upper panel) and burnout (lower panel) of the Peer Comparison and Leadership Training Condition (Condition 3) relative to the Peer Comparison Condition (Condition 2). Error bars reflect 95% confidence intervals.

$P = 0.013$, d = 0.39) than PCPs who received the peer comparison intervention alone (condition 2). These long-term effects on job satisfaction remained significant at the 5% level after a two-fold Holm–Bonferroni correction (condition 2 vs. 1: adjusted $P = 0.026$; conditions 3 vs. 2: adjusted $P = 0.026$). The long-term differences across conditions in burnout were not statistically significant, but they remained directionally consistent with the short-term treatment effects.

Together, these results indicate that the peer comparison intervention negatively impacted the following two dimensions of physician well-being: job satisfaction and burnout. The harmful effect on job satisfaction lasted for at least 4 mo after the intervention had been discontinued. However, administering the peer comparison intervention with leadership support training appeared to offset these harmful effects.

Exploratory Analysis of Mechanisms

Our finding that training leaders to be more supportive offset the negative effects of the peer comparison intervention on physician well-being led us to investigate one potentially important mechanism. We hypothesized that PCPs may have perceived the administration of the peer comparison intervention alone as signaling a lack of support from leadership (for instance, it may have seemed callous and misdirected). But adding leadership support training may have counteracted this impression. To test this hypothesis, we leveraged a measure of "perceived leadership support" that was included in our quarterly surveys ["I feel supported, understood, and valued by my department leaders" (35); 1-"strongly disagree" to 5-"strongly agree"].[†]

Figs. 3 and 4 depict the regression-estimated treatment effects of our interventions on perceived leadership support, based on the same regression specification that we used to predict job satisfaction and burnout (*SI Appendix*, Section 13). Compared to PCPs in the control condition (April 2020: M = 3.52, SD = 0.91; July 2020: M = 3.46, SD = 0.88), PCPs in the peer comparison condition (condition 2) reported feeling significantly less supported by their department leaders in both April 2020 (M = 3.02, SD = 1.21; β = −0.60, 95% CI = [−1.06, −0.13], $P = 0.013$, $d = 0.47$) and July 2020 (M = 2.87, SD = 1.21; β = −0.69, 95% CI = [−1.12, −0.26], $P = 0.002$, $d = 0.56$). However, PCPs who received leadership support training combined with peer comparison perceived significantly higher leadership support in April 2020 (condition 3: M = 3.55, SD = 1.06; β = 0.56, 95% CI = [0.09, 1.03], $P = 0.021$, $d = 0.47$) than PCPs who received the peer comparison intervention alone (condition 2). This difference is marginally significant in July 2020 (condition 3: M = 3.38, SD = 1.08; β = 0.49, 95% CI = [0.00, 0.98], $P = 0.054$, $d = 0.45$). Perceived leadership support did not significantly differ between condition 3 and the control condition in April 2020 ($P = 0.808$) or July 2020 ($P = 0.254$). Together, these results are consistent with the interpretation that the peer comparison intervention administered on its own caused PCPs to feel significantly less supported by their department leaders; but, importantly, leadership support training buffered against this effect.

To gain further insights into why the peer comparison intervention reduced perceived leadership support, we surveyed PCPs from our study population approximately 1 y after the intervention had ended (April 2021). Of the original 199 PCPs

[†]Who would be considered as "department leaders" was deliberately left open to the respondents' interpretation. For example, physician leads may have interpreted "department leaders" as referring to the health system's management, while nonlead physicians may have interpreted it as referring to their physician leads or nonclinical managers.

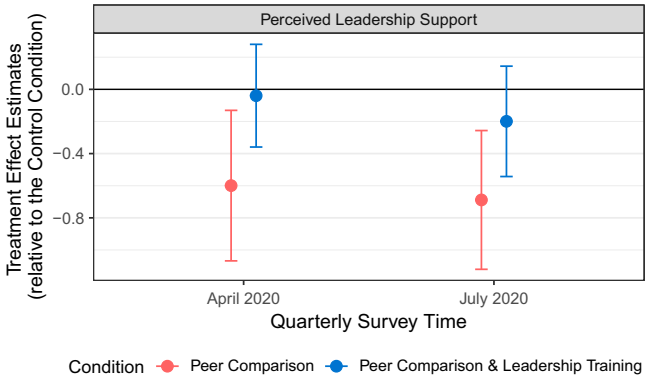

Fig. 3. Treatment effect estimates on perceived leadership support. The blue and red dots show the estimated treatment effects in the respective conditions (relative to the Control Condition) on perceived leadership support. The error bars reflect 95% confidence intervals.

in the experiment, 169 individuals (85%) were still working for UCLA and were thus invited to take the survey. Of these PCPs, 90.5% (153/169) completed part or all of the survey. Response rates did not significantly differ by the condition PCPs had been assigned to during our experiment ($P = 0.55$ for the F-test of joint significance).

In the survey, we first presented all PCPs (regardless of their experimental condition) with an example of the peer comparison email that had been used in our experiment, and we asked, "Would you prefer that the Department resumes sending these types of emails to physicians?" Of the 150 PCPs who responded to this question, 54% (81/150) preferred that the peer comparison emails not be resumed. More specifically, the proportion of PCPs preferring that peer comparison emails not be resumed was highest (i.e., 68%) among physicians from condition 2 who had experienced the peer comparison intervention alone (compared to 45% of PCPs from condition 1 and 50% from condition 3).

We next asked all PCPs an open-ended question about how receiving such peer comparison emails would make them feel. The open-ended responses again revealed PCPs' negative attitudes toward the peer comparison intervention (*SI Appendix*, Section 14 for more details). In particular, these responses suggested two related reasons why the peer comparison intervention would make PCPs feel less supported by leadership, and ultimately, less satisfied with their job and more burned out. First, the leadership's use of peer comparison information in this context was viewed by many PCPs as transgressive. For instance, one PCP stated, "frankly I think it is inappropriate"; another commented that "publicizing

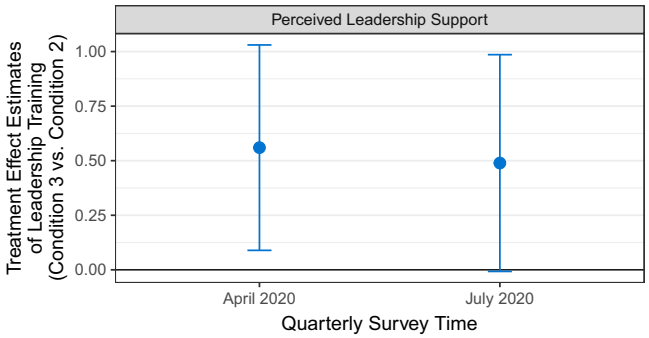

Fig. 4. Treatment effect estimates of leadership training on perceived leadership support. The blue dots show the estimated treatment effects of the Peer Comparison and Leadership Training Condition (Condition 3) relative to the Peer Comparison Condition (Condition 2). Error bars reflect 95% confidence intervals.

data among all faculty feels inappropriate, as if we are all being ranked/valued according to this metric." Second, leadership's use of one performance metric (the HM completion rate) in the peer comparison emails was viewed by many PCPs as too reductionist. For instance, one PCP stated that the HM completion rates "do not accurately gauge the quality of care a physician provide[s]"; another commented, "top physicians [are] defined by so much more than HM completion."

Several PCPs explicitly stated that they felt that the peer comparison emails should be accompanied by greater leadership and organizational support. For instance, one PCP cited a lack of "support from upper management to help"; another noted that "completion of health maintenance items should be a 'system' effort, not at the individual PCP level." The leadership support training provided participants (physician leads and nonclinical managers) with information on how completing HM measures would benefit patients, which they were encouraged to share with the nonparticipating PCPs in their clinics. We conjecture that such information may have helped PCPs—regardless of whether they participated in the training—contextualize the peer comparison emails, making them more amenable to accepting the HM completion metric as a marker of performance or showing them that management realized that this metric was not the only important measure of job performance. As a result, PCPs may have felt that their leaders were evaluating them more fairly and holistically when leadership support training was included as part of the intervention. Consistent with our speculation, leadership training in condition 3 did appear to improve the perceived leadership support both among physician leads who received the leadership training and among the nonlead physicians who did not personally attend the leadership training (*SI Appendix*, Table S25), although effects in these subgroups are no longer statistically significant due to the smaller sample sizes.

In summary, these qualitative responses suggest that the manner in which the peer comparison intervention was administered in our context was seen as normatively inappropriate and reductionist and that adding leadership support training buffered against these perceptions by helping leaders contextualize the intervention.

Discussion

Using a 5-mo field experiment involving 199 physicians and 46,631 patients, we examined the effects of a peer comparison intervention, administered alone or in conjunction with leadership support training, on physicians' preventive care performance and work-related well-being. In this setting, the peer comparison intervention did not significantly improve physicians' performance (measured as order rates for preventive measures). But it did unexpectedly harm job satisfaction and increase burnout, with the effect on job satisfaction persisting for at least 4 mo. Importantly, this negative effect of the peer comparison intervention on physician well-being was substantially attenuated by leadership support training. We find evidence that perceived leadership support may help explain both effects. The peer comparison intervention caused doctors to feel less supported by their leaders, but leadership training buffered against that negative effect.

Although we did not find a statistically significant effect of the peer comparison intervention on physician behavior, previous studies have found significant positive effects, even within similar contexts (6–8). Likewise, peer comparison interventions outside of the healthcare context have had inconsistent effects on targeted behaviors, with some showing null or negative effects (32, 33, 36–39) and others showing positive effects (1, 6, 8, 40–42). There are many different ways to operationalize and communicate peer

comparison interventions. We speculate that details in the intervention design, implementation, and context matter in determining their success. Among other aspects of our design, publicly displaying a list of the Top 25 performers using a composite performance metric may have curbed any motivating effects of peer comparisons for a few reasons. First, PCPs may have found it reductionist for their leaders to evaluate their job performance using a single metric (43). Second, it may have seemed unjust to evaluate performance in relative terms (i.e., Top 25), rather than using an absolute criterion that reflects the top quality of care (22). Using an absolute criterion instead would have also allowed for the public list of top performers to potentially grow over time, which could have motivated people by highlighting a growing trend (44). Third, highlighting exemplary performance (e.g., Top 25 physicians) could be discouraging to people who do not believe improvement is possible (12). In our case, people at the bottom of the performance distribution were the most likely to feel incapable of behavior change, even though they had the most room for improvement (*SI Appendix*, Section 15 for details). These features of our design may have been perceived as particularly inappropriate or offensive in the present social context, where physicians' roles and responsibilities typically involve communal norms that foster care and collaboration (5, 16).

Our findings offer three key contributions to the peer comparison literature. First, we provide field experimental evidence of the negative effects of a peer comparison intervention on workers' job satisfaction and burnout. Second, our findings underscore the importance of attending to the way in which implementation details of a peer comparison intervention are perceived by targeted individuals within the relevant social context. Researchers have recently argued that behavioral interventions are not experienced "in a vacuum," but rather that they are "embedded in a social ecosystem involving an implicit or explicit interaction between targeted individuals and the [intervention] designer" (45). According to this account, people attend to the details of behavioral interventions—especially interventions that have been newly introduced—to infer their leaders' beliefs and values. When such inferences are negative (e.g., my leaders do not seem to support me), targeted individuals may respond unfavorably to the intervention. Thus, to enhance the effectiveness of behavioral interventions, our research suggests that policymakers and organizational leaders ought to engage targeted individuals in the design phase of an intervention, probe the inferences they draw about it, and revise the design to reduce negative inferences before scaling the intervention in the field. Finally, our work highlights that when leaders offer the necessary context and support to accompany a peer comparison intervention, recipients may draw more positive inferences about their leaders' intent. This can buffer against the harmful effects of peer comparison interventions on well-being.

Our study has several limitations that suggest interesting directions for future research. First, our interventions had to be discontinued after only 5 mo due to the COVID-19 pandemic. It remains an open question whether the peer comparison intervention would have become normalized over time and thus might have stopped affecting physician well-being. Second, the leadership support training intervention was multifaceted with a variety of components and a broad curriculum. Future research is needed to discern which aspects of the leadership support training affected job satisfaction and burnout. Finally, although job satisfaction and burnout were preregistered secondary outcomes, we did not predict a negative effect a priori. It would be valuable to design future experiments to deductively test hypotheses concerning the conditions under which a broader range of behavioral interventions harm the well-being of targeted individuals.

When both the behavioral and psychological impact of an intervention are measured, difficult trade-offs may arise. How are we to decide whether an intervention is worthwhile if it produces desired behavior change (e.g., motivating physicians to improve patient outcomes) but reduces well-being? For instance, notifying doctors about their patients who suffered fatal overdoses has been shown to reduce subsequent opioid prescriptions (46). Although such notifications were likely highly aversive to doctors, one could argue that this is justified by the behavior change that saves lives. Naturally, other cases will be more ambiguous. In order to design and deploy interventions that holistically improve social welfare, researchers, policymakers, and ethicists will need to continue examining these trade-offs and develop approaches to quantify or even price the psychological consequences of interventions (9, 13).

Conclusion

Behavioral interventions such as providing peer comparison information offer attractive, cost-effective ways to promote positive behavior change. Our work suggests that if policymakers and organizational leaders only measure the behavioral outcomes of such interventions, they risk overlooking important effects on less visible outcomes, such as job satisfaction and burnout. These psychological outcomes need to be accounted for to estimate the aggregate impacts of policies and to improve their design and implementation.

Materials and Methods

Setting. Between November 5, 2019 and March 3, 2020, we collaborated with the UCLA Health Department of Medicine (DOM) Quality Team to run a field experiment across the health system's entire primary care network. In line with the DOM Quality Team's goal of motivating physicians to improve their patients' uptake of preventive care services, all PCPs in our study were part of a pay-for-performance program that incentivized them to meet a threshold HM completion rate. For each PCP, the HM completion rate reflects the proportion of recommended preventive care measures that were completed by their patients in a given time period. There are 26 different measures recommended by the US Preventive Service Task Force and other medical associations (e.g., American Diabetes Association), of which the DOM Quality Team identified nine high-priority "focus measures" (e.g., diabetes hemoglobin A1c screening). Details regarding how HM completion rates were calculated are available in *SI Appendix*, Section 1. This study was part of a quality improvement initiative implemented across the UCLA Health system and was determined to be exempt from review by the UCLA Institutional Review Board.

Experimental Design. The experiment was originally designed and preregistered to span 12 mo but was discontinued in March 2020 due to the COVID-19 pandemic (ClinicalTrials.gov no. NCT04237883). The experiment included 199 PCPs across 42 clinic sites that specialized in internal medicine, geriatrics, or family medicine and that had a clinical full-time employment rate of at least 50%. PCPs were unaware of this research investigation. They were cluster randomized at the clinic level to one of three study conditions, as follows: control, peer comparison, peer comparison and leadership training (Table 1). Each condition involved 14 clinics. For more information on the inclusion criteria and randomization algorithm, see *SI Appendix*, Sections 1 and 2. *SI Appendix*, Sections 3 shows that conditions were balanced on all observable patient, physician, and clinic characteristics.

All PCPs received monthly emails from the DOM Quality Team that informed them of their HM completion rate over the prior 3 mo. They were signed by the health system's management. The emails contained other information and links intended to help PCPs improve HM completion rates (*SI Appendix*, Sections 4 and 5 for email details, examples, and email engagement statistics). Emails were sent near the start of each month. A maximum of two reminder emails–identical to the initial email–were sent to those who had not opened the initial email after 7 and 14 d.

Table 1. Descriptions of intervention(s) implemented in each condition

Condition	Main intervention elements
1. Control	- Monthly emails informed PCPs of their HM completion rate over the prior 3 mo, the focus measure on which they had performed the best, and the two focus measures that they could most improve on
2. Peer comparison	- Same information as in the monthly emails in the Control Condition - Monthly emails also included a list of the names of the top 25 PCPs as well as messaging based on the recipient's placement in the performance distribution (Top 25 Physician, High Performer, Almost High Performer, Low Performer)
3. Peer comparison and leadership training	- Same monthly emails as in the peer comparison condition - Clinical physician leaders and nonclinical managers received two training workshops (on how to provide effective support to fellow physicians) and monthly check-in emails - Physician leads had one-on-one meetings with members of the DOM Quality Team to identify specific challenges at their clinics and brainstorm strategies to address these challenges

For PCPs in the peer comparison condition (condition 2) and the peer comparison and leadership training condition (condition 3), the emails also included information about their peers' performance. These emails contained a banner displaying the names of PCPs whose HM completion rate in the prior 3 mo was within the top 25 of all PCPs in the study population. These PCPs were labeled Top 25 Primary Care Physicians. Additionally, emails in conditions 2 and 3 informed PCPs of their relative standing in terms of HM completion rates compared to all other PCPs in the prior 3 mo.

- PCPs who were one of the Top 25 PCPs in a given month received a message saying, "Congratulations! You are a Top 25 Primary Care Physician in [respective month]!"
- PCPs whose HM completion rate was above 65% but who were not one of that month's Top 25 PCPs were informed, "Congratulations! You are a High Performer!"
- PCPs whose HM completion rate was between 55 to 65% were told, "You are almost a High Performer."
- PCPs with an HM completion rate under 55% were informed that "the majority of physicians have a HM completion rate of 55% or higher."

The emails further informed all PCPs of the HM completion rates necessary to be a "High Performer" or a "Top 25 Primary Care Physician," whichever was more proximate, and they encouraged PCPs to improve their performance (or maintain their performance if they were already a Top 25 PCP). The performance tier cutoffs had been selected to ensure that most PCPs would fall into the two middle performance tiers, where they would feel close to reaching the next-higher group (*SI Appendix*, Section 4 for details).

For the 14 clinics assigned to the peer comparison and leadership training condition (condition 3), physician leads and nonclinical managers participated in two 4-h training workshops, namely, one in December 2019 and one in March 2020. These workshops focused on training attendees to develop their leadership skills and effectively support their fellow PCPs. Importantly, a primary goal of the training was to help attendees provide fellow PCPs at their clinics with the necessary contextual information to understand and appreciate why UCLA Health uses HM completion rates to measure performance. Among the physician leads in condition 3 clinics, 11 were in our experiment, overseeing a total of 59 other PCPs.

Following the training workshops, clinic physician leads and their nonclinical manager counterparts received additional resources from the DOM Quality Team through monthly check-in emails. Additionally, in-person meetings with clinic physician leads in condition 3 occurred in January and February 2020. During these one-on-one meetings, the DOM Quality Team helped physician leads identify specific challenges at their clinic and develop corresponding solutions. *SI Appendix*, Section 6 includes detailed information about the leadership training intervention along with materials from the training workshops.

Data. For each PCP, we measured their order rate for patients who satisfy the following preregistered criteria: 1) patients were empaneled to a PCP participating in the field experiment (based on the attribution logic laid out in *SI Appendix*, Section 1), 2) had at least one in-office visit with their PCP during the intervention period (November 5, 2019 to March 3, 2020), and 3) had at least one focus measure due at the time of that in-office visit. A total of 46,631 patients met these inclusion criteria. See *SI Appendix*, Section 3 for more information on the sample characteristics.

From October 2019 through July 2020, PCPs who were part of the field experiment were asked to complete quarterly surveys assessing their experiences at work and participation in professional activities. The surveys (sent by the DOM leadership) collected longitudinal measures of job satisfaction, burnout, and feelings of leadership support, along with other measures not pertinent to the current investigation. Since the field experiment had to be discontinued in March 2020 (due to COVID-19), we used the April 2020 survey data for our primary analysis. We also examined the sustained impact of our interventions by analyzing the July 2020 survey data. The completion of these surveys was tied to the aforementioned pay-for-performance incentive program. Thus, 93.0% (185/199) of physicians completed the April 2020 survey and 88.4% (176/199) completed the July 2020 survey (*SI Appendix*, Section 7 for survey details and quarterly completion rates). See Fig. 5 for a timeline of the study.

Measures. For each patient empaneled to a PCP, our preregistered primary behavioral outcome was the HM order rate for focus measures that were due at the patient's first in-office primary care visit during the intervention period (hereafter, order rate). It equals the share of open HM focus measures (i.e., focus measures recommended for the patient based on the national guidelines but not yet completed at the time of the patient's first visit) that were ordered by the PCP within 7 d following the patient's first visit:

$$\text{Order Rate} = \frac{\text{Number of ordered HM focus measures 7 d following first visit during the study period}}{\text{Number of open HM focus measures at the time of the first visit during the study period}}.$$

The order rate was chosen as the primary behavioral outcome because it is clinically important and not subject to factors outside the PCPs' control (e.g., patients' willingness or ability to obtain preventive service).

Our preregistered secondary outcomes included two measures of physician well-being, namely, job satisfaction and burnout, which we assessed using validated single-item scales in every quarterly survey. Job satisfaction was measured with the question, "Taking everything into consideration, how do you feel about your job as a whole?", with responses ranging from "extremely dissatisfied" to "extremely satisfied" on a seven-point Likert scale (47). We used a validated and widely used burnout measure (48), as follows:

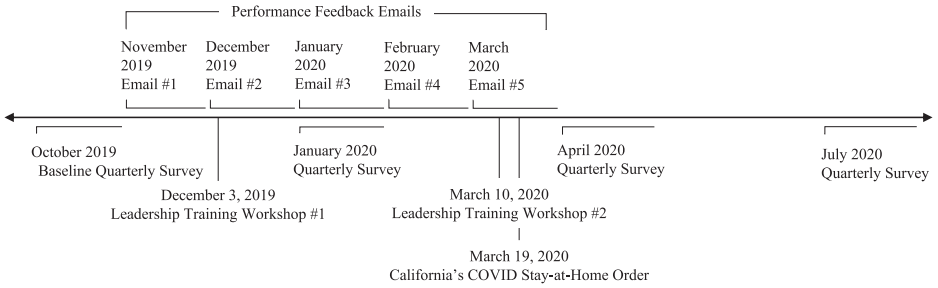

Fig. 5. Study timeline. This timeline depicts the timing of the relevant events in the study. The L-shaped lines depict events that occurred over sustained periods of time. The performance feedback emails were initially sent at the beginning of each month, and up to two reminders were sent during the month to those who had not opened the initial emails; PCPs had approximately 2 wk to complete the quarterly surveys. The straight vertical lines depict discrete events. For ease of visualization, the email and survey dates are approximate. See *SI Appendix*, Section 4 for the precise dates of each email sent and survey launched.

"Overall, based on your definition of burnout, how would you rate your level of burnout?", with five response options ranging from 1 = "I enjoy my work. I have no symptoms of burnout" to 5 = "I feel completely burned out and often wonder if I can go on. I am at the point where I may need some changes or may need to seek some sort of help."

Statistical Analysis. To compare patient-level order rates between conditions, we estimated a mixed effects binomial logistic regression model. This model assumes that each patient's number of orders placed follows a binomial distribution, where the number of trials is the patient's number of open topics and a logit-linear function is used to estimate the probability that a patient has an order placed for any given open topic. Physician and clinic random effects account for clustering of patients. The preregistered baseline controls are as follows: patient characteristics, including their completion rate measured from July to October 2019, age, gender, and zip code (using fixed effects for the three-digit zip code for all Southern California zip codes and a single indicator for everyone else),[‡] and physician characteristics, including their gender, race, years since graduating medical school, and years of working at UCLA Health. We preregistered the following gatekeeping approach for our analysis in order to reduce multiple hypothesis testing (49): we would first test whether HM order rates differed between the combination of conditions 2 and 3 versus condition 1. If and only if this comparison was statistically significant, we would conduct additional comparisons across conditions using a Holm–Bonferroni adjustment, for an overall significance level of 0.05. Our results are robust to alternate specifications including binomial logistic regressions with SEs clustered at the clinic level, mixed effects linear models with physician and clinic random effects, and linear regression models with SEs clustered at the clinic level (reported in *SI Appendix*, Section 8).

To assess differences in survey measures (e.g., job satisfaction, burnout) between conditions, we used linear regression models, with cluster-robust SEs at the clinic level. These regressions controlled for the respective outcome measure taken from the baseline October 2019 quarterly survey, as well as the same set of physician demographics as preregistered in our analysis of order rates (physician gender, race, years since graduating medical school, and years of working at UCLA Health).

Data Availability. The code to replicate the analyses and figures in the article and the *SI Appendix* has been deposited in ResearchBox (https://researchbox.org/654) (50). The data analyzed in this article were provided by UCLA Health and may contain protected health information. To protect participant and patient privacy, we cannot publicly post individual-level data. Qualified researchers with a valuable research question and relevant approvals including ethical approval can request access to the deidentified data about these trials from the corresponding author. A formal contract will be signed and an independent data protection agency should oversee the sharing process to ensure the safety of the data.

ACKNOWLEDGMENTS. We thank UCLA Health Department of Medicine for allowing us to conduct this intervention and assisting with data collection. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant no. DGE-1650604. We thank Jose Cervantez for his outstanding research assistance.

[‡]We had also preregistered controlling for patient comorbidity and insurance plan. We unexpectedly did not have access to these variables, and thus, they are not included in the reported regressions.

Author affiliations: [a]University of California, Los Angeles (UCLA) Anderson School of Management, Los Angeles, CA 90095; [b]David Geffen School of Medicine at UCLA, Los Angeles, CA 90095; [c]UCLA Health Department of Medicine, Los Angeles, CA 90095; and [d]Central Michigan University, Mt. Pleasant, MI 48859

1. A. Tran, R. Zeckhauser, Rank as an inherent incentive: Evidence from a field experiment. *J. Public Econ.* **96**, 645–650 (2012).
2. H. Allcott, T. Rogers, The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *Am. Econ. Rev.* **104**, 3003–3037 (2014).
3. A. S. Gerber, T. Rogers, Descriptive social norms and motivation to vote: Everybody's voting and so should you. *J. Polit.* **71**, 178–191 (2009).
4. B. S. Frey, S. Meier, Social comparisons and pro-social behavior: Testing conditional cooperation in a field experiment. *Am. Econ. Rev.* **94**, 1717–1722 (2004).
5. S. Blader, C. Gartenberg, A. Prat, The contingent effect of management practices. *Rev. Econ. Stud.* **87**, 721–749 (2020).
6. D. Meeker *et al.*, Effect of behavioral interventions on inappropriate antibiotic prescribing among primary care practices. *JAMA* **315**, 562–570 (2016).
7. H. Song, A. L. Tucker, K. L. Murrell, D. R. Vinsonc, Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Manage. Sci.* **64**, 2628–2649 (2018).
8. A. S. Navathe *et al.*, Assessing the effectiveness of peer comparisons as a way to improve health care quality. *Health Aff. (Millwood)* **39**, 852–861 (2020).
9. H. Allcott, J. B. Kessler, The welfare effects of nudges: A case study of energy use social comparisons. *Am. Econ. J. Appl. Econ.* **11**, 236–276 (2019).
10. D. J. Brown, D. L. Ferris, D. Heller, L. M. Keeping, Antecedents and consequences of the frequency of upward and downward social comparisons at work. *Organ. Behav. Hum. Decis. Process.* **102**, 59–75 (2007).
11. T. Rogers, A. Feller, Discouraged by peer excellence: Exposure to exemplary peer performance causes quitting. *Psychol. Sci.* **27**, 365–374 (2016).
12. P. Lockwood, Z. Kunda, Superstars and me: Predicting the impact of role models on the self. *J. Pers. Soc. Psychol.* **73**, 91–103 (1997).
13. L. Butera, R. Metcalfe, W. Morrison, D. Taubinsky, Measuring the welfare effects of shame and pride. *Am. Econ. Rev.* **112**, 122–168 (2022).
14. H. Hermes, M. Huschens, F. Rothlauf, D. Schunk, Motivating low-achievers–Relative performance feedback in primary schools. *J. Econ. Behav. Organ.* **187**, 45–59 (2021).
15. J. M. T. Krijnen, D. Tannenbaum, C. R. Fox, Choice architecture 2.0: Behavioral policy as an implicit social interaction. *Behav. Sci. Policy* **3**, i–18 (2017).
16. J. Gallus, J. Reiff, E. Kamenica, A. P. Fiske, Relational incentives theory. *Psychological Review* **129**, 586–602 (2022).
17. A. Bobbio, M. Bellan, A. M. Manganelli, Empowering leadership, perceived organizational support, trust, and job burnout for nurses: A study in an Italian general hospital. *Health Care Manage. Rev.* **37**, 77–87 (2012).
18. C. Maslach, W. B. Schaufeli, M. P. Leiter, Job burnout. *Annu. Rev. Psychol.* **52**, 397–422 (2001).
19. T. D. Shanafelt *et al.*, Impact of organizational leadership on physician burnout and satisfaction. *Mayo Clin. Proc.* **90**, 432–440 (2015).
20. C. P. West, L. N. Dyrbye, T. D. Shanafelt, Physician burnout: Contributors, consequences and solutions. *J. Intern. Med.* **283**, 516–529 (2018).
21. A. McKethan, A. K. Jha, Designing smarter pay-for-performance programs. *JAMA* **312**, 2617–2618 (2014).
22. T. Mayer, A. Venkatesh, D. M. Berwick, Criterion-based measurements of patient experience in health care. *JAMA* **326**, 2471–2472 (2021).
23. A. Hassol, N. West, J. Gerteis, M. Michaels, "Evaluation of the oncology care model" (Abt Associates, Rockville, MD, 2021).
24. *Care Compare, Doctors and Clinicians Initiative*. (Centers for Medicare and Medicaid Services, 2021).
25. L. Kane, "Medscape national physician burnout & suicide report" (Medscape, 2021).
26. J. M. George, G. R. Jones, The experience of work and turnover intentions: Interactive effects of value attainment, job satisfaction, and positive mood. *J. Appl. Psychol.* **81**, 318–325 (1996).

27. T. A. Judge, C. J. Thoresen, J. E. Bono, G. K. Patton, The job satisfaction-job performance relationship: A qualitative and quantitative review. *Psychol. Bull.* **127**, 376–407 (2001).
28. B. M. Staw, R. I. Sutton, L. H. Pelled, Employee positive emotion and favorable outcomes at the workplace. *Organ. Sci.* **5**, 51–71 (1994).
29. T. A. Wright, D. G. Bonett, The contribution of burnout to work performance. *J. Organ. Behav.* **18**, 491–499 (1997).
30. S. Han *et al.*, Estimating the attributable cost of physician burnout in the United States. *Ann. Intern. Med.* **170**, 784–790 (2019).
31. S. W. Yates, Physician stress and burnout. *Am. J. Med.* **133**, 160–164 (2020).
32. N. Ashraf, O. Bandiera, S. S. Lee, Awards unbundled: Evidence from a natural field experiment. *J. Econ. Behav. Organ.* **100**, 44–63 (2014).
33. O. Bandiera, I. Barankay, I. Rasul, Team incentives: Evidence from a firm level experiment. *J. Eur. Econ. Assoc.* **11**, 1079–1114 (2013).
34. J. E. Bogard, M. A. Delmas, N. J. Goldstein, I. S. Vezich, Target, distance, and valence: Unpacking the effects of normative feedback. *Organ. Behav. Hum. Decis. Process.* **161**, 61–73 (2020).
35. S. F. Richer, R. J. Vallerand, Construction and validation of the ESAS (The relatedness feelings scale). *Eur. Rev. Appl. Psychol.* **48**, 129–138 (1998).
36. H. Hennig-Schmidt, A. Sadrieh, B. Rockenbach, In search of workers' real effort reciprocity-a field and a laboratory experiment. *J. Eur. Econ. Assoc.* **8**, 817–837 (2010).
37. I. Barankay, "Rank incentives: Evidence from a randomized workplace experiment" (Business Economics and Public Policy Papers, Wharton School of the University of Pennsylvania, Philadelphia, PA, 2012).
38. F. Buntinx *et al.*, Does feedback improve the quality of cervical smears? A randomized controlled trial. *Br. J. Gen. Pract.* **43**, 194–198 (1993).
39. L. Bursztyn, R. Jensen, How does peer pressure affect educational investments? *Q. J. Econ.* **130**, 1329–1367 (2015).
40. G. Azmat, N. Iriberri, The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *J. Public Econ.* **94**, 435–452 (2010).
41. J. B. I. Vidal, M. Nossol, Tournaments without prizes: Evidence from personnel records. *Manage. Sci.* **57**, 1721–1736 (2011).
42. W. Verbeke, R. P. Bagozzi, F. D. Belschak, The role of status and leadership style in sales contests: A natural field experiment. *J. Bus. Res.* **69**, 4112–4120 (2016).
43. A. Ranganathan, A. Benson, A numbers game: Quantification of work, auto-gamification, and worker productivity. *Am. Sociol. Rev.* **85**, 573–609 (2020).
44. G. Sparkman, G. M. Walton, Dynamic norms promote sustainable behavior, even if it is counternormative. *Psychol. Sci.* **28**, 1663–1674 (2017).
45. C. R. Fox *et al.*, Details matter: Predicting when nudging clinicians will succeed or fail. *BMJ* **370**, m3256 (2020).
46. J. N. Doctor *et al.*, Opioid prescribing decreases after learning of a patient's fatal overdose. *Science* **361**, 588–590 (2018).
47. C. L. Dolbier, J. A. Webster, K. T. McCalister, M. W. Mallon, M. A. Steinhardt, Reliability and validity of a single-item measure of job satisfaction. *Am. J. Health Promot.* **19**, 194–198 (2005).
48. E. D. Dolan *et al.*, Using a single item to measure burnout in primary care staff: A psychometric evaluation. *J. Gen. Intern. Med.* **30**, 582–587 (2015).
49. A. Dmitrienko, A. C. Tamhane, Gatekeeping procedures with clinical trial applications. *Pharm. Stat.* **6**, 171–180 (2007).
50. J. S. Reiff *et al.*, Data from "When peer comparison information harms physician well-being." ResearchBox. https://researchbox.org/654. Deposited 14 April 2022.