Private and Public Performance Reports as Drivers of Performance and Determinants of Performance Measure Information Content

A thesis presented

by

Henry Christian Eyring

In partial fulfillment of the requirements
for the degree of
Doctor of Business Administration

Harvard University
Graduate School of Business Administration
Cambridge, Massachusetts

April 2017

Professor Dennis Campbell **Henry Eyring**

Private and Public Performance Reports as Drivers of Performance and Determinants of Performance Measure Information Content

Abstract

This dissertation addresses how private and public performance reports affect performance and the information content of performance measures. First, I show how disclosing consumer ratings to the general public affects performance and biases raters. Using data from a health care system, I find that publicly disclosing patient ratings of physicians leads to: 1) performance improvement by the ratings and by objective measures of quality, and 2) a bias among raters, who positively weight a physician's published average rating in deriving subsequent ratings for the physician. To understand the moderating effects of public attention, I use variation in web traffic to a physician's disclosed rating. I find evidence consistent with public attention reinforcing raters' bias toward concurring with a physician's published average rating, thus impeding rating improvement. Within a national distribution of ratings, the disclosure leads to an improvement in ratings by 17 percentile points and a bias in a given physician's ratings toward his or her published average rating by 24 percentile points. These findings demonstrate that consumer-rating disclosure is a means of performance management, and that resulting bias is a reason to interpret subsequent trends in ratings as understated signals of trends in service.

The second section of the dissertation shows an understudied and low-cost way of customizing private performance reports to best drive reported performance, and warns that the private reporting causes reported performance to diverge from unreported performance. A field experiment reveals the performance benefit of customizing a private performance report to include the peer-performance reference point that will most motivate improvement. The below-

average performers improve most when shown the median as a reference point. The 50^{th}-75^{th} percentile performers improve most when shown the top-quartile as a reference point. The top-quartile performers improve most when shown the top-quartile as a reference point, but only when reported performance is outcome-based as opposed to process-based. Neither the median nor top-quartile reference point has a more positive performance effect overall. With regard to the performance measure's content, privately reporting a measure causes the measure to become less correlated with unreported performance. These findings have the following implications. First, the optimal reference point for peer performance comparison depends on 1) an individual's initial performance relative to each reference point, and 2) whether the performance measure regards an outcome or process. Second, a performance measure, once reported, becomes a less informative signal of unreported performance.

Acknowledgements

I am profoundly grateful for my mentors Dennis Campbell, V.G. Narayanan, Srikant Datar, and Ananth Raman. I have never associated with finer minds or more dedicated advocates. They have honed my rough ideas to inform challenges in the effective application of accounting and management. They have offered patient, detailed instruction, and have gone to extraordinary lengths in providing me with opportunities. I have often watched a mentor go out on a limb to persuade organizations and colleagues to provide access to resources for research. In asserting my value, they both risked their reputations and provided a model for mentorship. They saw me as someone I had yet to become, but that they were sure I could, and then guided me amid the inevitable starts and stops along the way.

I drew my writing style from Dennis Campbell, and have walked through each sentence of dissertation chapters with his guidance. He has methodically provided me with traction in developing as a researcher since I arrived in the program. V.G. has treated me like a son. He has carried me on his shoulders at points in navigating interactions with field sites, journals, and colleagues. I could go to him for anything, and I feel that I have. Srikant Datar has introduced me to many executives in health care, providing me access to the majority of the data used in my research. He also connected me to Bob Kaplan, who paved a relation that afforded the data for my job market paper. Srikant and Bob are renowned among business executives in part because of the applicability of their scholarship, and they have each met with me many times to help me target my research toward the most challenging dilemmas in health care. Ananth Raman has left me with a solid vision of what I want to do as an academic and as a provider for my family. That comes in part through his direct advice, which he has generously

offered since my first day of class at Harvard Business School, and just as much through observing him.

Numerous organizations have invested heavily in providing me access to data and the ability to test interventions through field experiments. Vivian Lee and her colleagues at University of Utah Health Care, and Heather Sternshein and her colleagues at HarvardX, are foremost among these. Each organization has served as an institution of my education to as great a degree as has Harvard. They have extended the greatest patience in allowing me to learn and conduct interventions and analyses. They are angel investors in my research and I would not have any of the results in this dissertation or nearly any other projects without their willingness to give me multiple chances.

I have relied heavily on the guidance of Kim Clark, Clark Gilbert, and Clayton Christensen since before my program began. They directed and contributed to my preparation for the program to an immeasurable degree. They have also provided invaluable counsel on developing a pipeline that forms a cohesive line of inquiry.

My sincerest thanks go to my father, mother, and my wife. If my father were a DBA, he would have accomplished more than me. My work is in many regards a watered-down version of his insight put in writing. We speak by phone for hours every week. He has treated me as a peer and equal since I was a toddler, and so we never had to transition to being colleagues. My mother is first-rate in her interpersonal skills. I try to treat people the way that she does and taught me to. That has formed a basis for friendships and collaborations in the program. From a distance and on visits she has carefully tended to my temporal and spiritual needs. My wife has walked each step of the way with me, through hills and valleys. When I am sleepless, she is too. When something works out well for me, she seems happier than I am. She has vicarious relationships

with all of my colleagues through me. I will strive to be as selflessly invested in her efforts to fulfill her dreams as she is in mine.

Table of Contents

CHAPTER 1

INTRODUCTION

Though health care costs are soaring, so is the availability of cost measures for tracking and managing costs. Though structural gaps between job growth and unemployed workers' skillsets are expanding, so is the availability of student behavior measures for tracking, customizing, and improving performance toward acquiring a skill. In education, health care, and other industries with vexing problems, then, the value of aptly applied measurement has perhaps never been greater. As "specialists of measurement," accounting scholars have an instrumental role to play in solving the great economic conundrums of our time (Van der Stede [2015]).

Measurement, though, is not a one-dimensional management tool that can simply be ramped up to meet management problems in the same way that supply can be ramped up to meet demand. Measures can be ill suited to a goal, backfiring by misdirecting effort. They can even discourage effort at all, or worse, lead to cheating. Sad examples span from multinational fraud, like those at Toshiba and Wells Fargo, to cheating in school districts like Atlanta and Washington, D.C.

Those examples punctuate the need for informed measurement. This dissertation is an effort to inform the producers and users of measurement systems to get the most upside and the least downside. It begins with an analysis in Chapter 2 of physician rating disclosure. I look at performance effects as well as a cost in terms of information content; performance by reported and unreported performance improves, but the subjective ratings once disclosed become sticky around their previously published values. This suggests a reason for companies to disclose their employees' customer ratings, as well as a reason to then interpret changes in

ratings as understated due to their stickiness toward past values. Without reporting the ratings, organizations pass up a low-cost opportunity to drive performance. Without updating the interpretation of ratings, organizations and consumers lose an opportunity to identify and reward changes in performance.

Chapter 3 draws on a study with V.G. Narayanan, and shows a performance benefit of privately disclosing to online education students their performance relative to standards of peer performance. Though a number of studies have found that displaying such reference points drives performance, little to no research informs which reference point to display. We provide such evidence, finding that the optimal performance reference point to display to an individual depends on their initial performance relative to that reference point. We also find a cost in terms of information content. The reported measure becomes less correlated with unreported measures of performance. This is a cost in terms of performance inference—the reported measure can no longer be seen as so indicative of unreported, and perhaps unmeasured, performance. It is also a cost if performance by the reported measure is desirable when it comes along with important related performance—such as a grade coming with thorough engagement in the course, which we find is less common after privately reporting to individuals their grades relative to a peer standard. We thereby offer guidance for organizations to maximize performance by a reported measure through providing the appropriate reference point to each individual. We also raise the caveat the organizations should then monitor unreported performance to make sure that it does not lag behind reported performance in a way that undermines the value of improvement by reported performance.

CHAPTER 2
DISCLOSING PHYSICIAN RATINGS: PERFORMANCE EFFECTS AND THE DIFFICULTY OF ALTERING RATINGS CONSENSUS

2.1 INTRODUCTION

This study investigates effects of a health care system disclosing patient ratings of its physicians to the public. Health care systems, hotel chains, and universities are among the many other organizations disclosing consumer ratings.[1] Though studies find disclosure of other types driving performance, consumer-rating disclosure is distinct in that it reveals subjective ratings to subsequent raters.[2] Behavioral economic mechanisms free to operate under those conditions may bias ratings toward the published consensus (Furnham and Boo [2011], Tversky and Kahneman [1975]), which would dull the sensitivity of ratings to effort and hamper their improvement (Banker and Datar [1989]). Each end of the resulting theoretical tradeoff, whereby consumer-rating disclosure may elicit improvement despite weighting consumer ratings toward the published consensus, is relevant to research on performance disclosure and on the effective use of consumer ratings.[3] Evidence regarding this tradeoff, though, is lacking.

I assess the predicted tradeoff and its dynamics empirically using data from the disclosure of physician ratings at University of Utah Health Care (UUHC).[4] UUHC offered research access to visit-level data from millions of patient visits occurring over more than three-and-a-half years. The tests herein exploit variation in whether and when physicians became subject to the

[1] For examples of consumer-rating disclosure by such institutions, see Cleveland Clinic [2016], Stanford Health Care [2016], Starwood [2016], Marriott [2016], Holiday Inn [2016], Columbia University [2016], and Texas Tech University [2016].

[2] Bennear and Olmstead [2008], Jin and Leslie [2003], Lu [2012], and Chatterji and Toffel [2010] address public performance disclosure by third-party evaluators, and do not indicate that the disclosure alters the information available to evaluators. Further, with the exception of Chatterji and Toffel [2010], the disclosed measures are objective in nature (though Jin and Leslie [2003] address restaurant hygiene grades, they state that the grades' "subjective component has been removed" since before effect estimation).

[3] See Leuz and Wysocki [2016] for a survey of performance disclosure literature, and Kaplan and Norton [2005] and Luca [2016] regarding uses of consumer ratings.

[4] "Physician ratings" is one of the popular terms used to describe patient ratings of physicians (e.g., Glover [2014]), and I adopt this terminology.

disclosure. A generalized difference-in-differences approach pools estimates from the disclosure intervention's staggered implementation. Such pooling, used in prior research on health care disclosure, increases the estimates' precision and robustness (Dranove, Kessler, McClellan, and Satterthwaite [2003], Duflo [2002]). The sample's substantial time range allows validating the assumption of parallel dependent-variable trends. Physician fixed-effects control for static differences among physicians included in and excluded from the disclosure. Robustness tests, including propensity-score matching, suggest that demographic differences among physicians do not drive results. Data on patient characteristics, including multiple measures of underlying patient health, allow controlling for patient mix through means consistent with leading health economics research (e.g., Chandra, Gruber, and McKnight [2010], Dafny [2005], Doyle Jr. [2011]). Isolating changes in performance from changes in patient mix is critical in assessing health care disclosure's performance effects (Dranove et al. [2003]). With this identification strategy, I address the noted theoretical tradeoff of consumer-rating disclosure.

On one end of the tradeoff, consumer-rating disclosure may elicit performance improvement. This result is not theoretically straightforward. Bias among raters after viewing disclosed ratings could dull the sensitivity of ratings as performance measures, deterring and/or misdirecting service providers' effort toward rating improvement. Regarding effort deterrence, the economically optimal level of effort to exert toward improving a measure declines with a reduced sensitivity of the measure to effort (Banker and Datar [1989]). Regarding effort misdirection, rater bias would obscure responses of ratings to fundamental changes in service, inhibiting physicians' ability to learn through trial-and-error (Campbell, Epstein, and Martinez-Jerez [2011]). Extant disclosure literature shows performance effects in settings wherein a bias in performance measures toward published values, and related performance-impeding forces, are

not mentioned and are relatively unlikely.[5] The current study is unique in testing consumer-rating disclosure's performance effects, and establishing their persistence among ratings' bias toward the published consensus.

A bias in consumer ratings toward the published consensus rating constitutes the other end of the predicted tradeoff. I term this a "consensus-bias" effect of disclosing consumer-ratings, and provide the first substantiating evidence.[6] Raters subject to consensus bias would positively weight the published consensus rating in forming their own ratings. The effect may result from the anchoring and adjusting heuristic among raters, whereby an initially displayed reference point attracts subsequent estimates toward itself. However, anchoring has primarily been established under the conditions of an *arbitrary* reference point and estimation regarding a topic that an individual has *limited* familiarity with (Furnham and Boo [2011], Tversky and Kahneman [1975]). Whether anchoring applies when *informative* reference points – actual prior consumer ratings – are available to an individual who has *certainly* experienced the subject they are evaluating is unclear. The relevance of prior consumer ratings would plausibly make them more salient, and thereby influential, as subconscious reference points. Direct and recent experience with the subject of rating, though, may facilitate forming one's independent rating (Muchnik et al. [2013]). In addition to anchoring, social herding may yield consensus bias. However, related theory hinges on individuals feeling uncertain and assuming that others are

[5] Bennear and Olmstead [2008] assess disclosure of water safety violations, Jin and Leslie [2003] restaurant hygiene grades whose "subjective component has been removed" since before effect estimation, and Lu [2012] percentages of nursing home patrons with various health problems. The objective nature of these measures counters the use of subjectivity among evaluators to produce biased evaluations. Further, in each of those studies and in Chatterji and Toffel [2010], regarding subjective rating disclosure, the disclosures are from third-party evaluators to the public. The studies do not mention the disclosure altering the information available to evaluators, nor potentially biasing performance evaluations toward published values.

[6] A study of Amazon reviews estimates unbiased scores, but does not posit or assess whether bias is toward or away from the published consensus (Sikora and Chuahan [2012]), a lab study finds mixed evidence that consumers who disagree with a consensus rating amplify their rating in the direction of disagreement (Eryarsoy and Piramuthu [2014]), and a third study shows a net effect of an *arbitrarily* assigned "thumbs-up," but not of a "thumbs-down," on average ratings of web content (Muchnik, Aral, and Taylor [2013]).

better informed, which may not transfer to the case of a direct consumer evaluating their consumption experience (Baddeley [2010], Keynes [1930]). Further, theories of consumer behavior suggest that some consumers who disagree with published rating exaggerate their ratings in the direction of disagreement, and a lab study finds mixed evidence of this (Eryarsoy and Piramuthu [2014]).

The current study extends three streams of research. The first is research on performance effects of disclosure (e.g., Bennear and Olmstead [2008], Chatterji and Toffel [2010], Jin and Leslie [2013]). I extend this stream to the growing realm of consumer-rating disclosure. I find that consumer-rating disclosure is an effective tool for lifting performance, and that it must overcome consensus bias among raters in that process. The performance effects include improvement by consumer ratings and by objective quality measures. Improvement by those nonfinancial performance measures has been shown to drive financial performance, and also to be difficult to achieve through financial contracts.[7, 8] I provide the first evidence of the tradeoffs of an alternative approach to incentivizing nonfinancial performance – that of disclosing consumer ratings.

Second, this paper speaks particularly to the stream of economic and medical research on the performance effects of health care disclosure. Extant literature in this stream has focused on outcome and safety measures, and has found limited performance effects.[9] Physician ratings reportedly garner greater interest among patients than do these other types of health care

[7] Ittner, Larcker, and Meyer [2003] report that supervisors tasked with administering balanced-scorecard-based compensation shifted weight away from nonfinancial performance measures, and Bol, Keune, Matsumura, and Shin [2010] report that supervisors are adversely influenced by political concerns in applying nonfinancial performance ratings within compensation.

[8] See Chevalier and Mayzlin [2006], Hu, Liu, and Zhang [2008], and Luca [2016] regarding effects of online reviews on revenue, Ittner and Larcker [1998], Banker et al. [2000], and Nagar and Rajan [2005] regarding effects of customer satisfaction on financial performance, and Balasubramanian, Mathur, and Thakur [2005], and Banker et al. [2000] regarding effects of quality on financial performance.

[9] See Dranove et al. [2003], Epstein [2006], Ryan, Nallamothu, and Dimick [2012], and Shukla [2013] finding small, inconclusive, or no effects of such health care disclosure initiatives.

performance (Brown, Clarke, and Oakley [2012], Dafny and Dranove [2008], Hanauer, Zheng, Singer, Gebremariam, and Davis [2014]). Public attention to disclosure should, in theory, facilitate the disclosure's performance effects (Kolstad [2013], Parker and Nielsen [2011], Weil et al. [2006]), which raises the possibility that disclosing physician ratings will yield significant performance effects. This paper's assessment of those performance effects is pertinent in light of the many health care systems, including industry leaders such as Stanford Health Care and Cleveland Clinic, that have recently disclosed physician ratings.

This paper extends a third stream of literature, on consumer ratings. Consensus bias would delay ratings in depicting a new service level while the published consensus rating, toward which the ratings would be biased, updated to reflect the new service level. Managers and researchers aware of such a lag in ratings depicting a new service level could better assess the sequence of service improvement, online consumer ratings, and financial performance. Accounting and economic studies note the importance of this sequence to managers (Banker et al. [2000], Banker and Mashruwala [2009], Luca [2016]). Also, principals could account for a lag in ratings' responsiveness to service-level changes in using the ratings to infer agents' effort (Lyu, Wick, Housman, Freischlag, and Makary [2013], Ubel [2015]). Such inference is a step in mitigating moral hazard (Hölmstrom [1979]). Finally, consensus bias is relevant to literature on employee evaluation biases in two regards (Bol [2011], Moers [2005], Prendergast and Topel [1993]). First, consensus bias among consumer raters would affect employee evaluations that incorporate consumer ratings. Second, supervisors who are aware of a past evaluation consensus regarding an employee may be prone to consensus bias in their own evaluations of the employee (Grote [2005]).

To further advance the noted streams of research, I explore the way in which public attention to disclosure moderates its performance and bias effects. Using data on web traffic to individual physicians' disclosed ratings I assess evidence of theory that public attention to disclosure strengthens its performance effects (Kolstad [2013], Parker and Nielsen [2011], Weil et al. [2006]). Further, I show how consumer-rating improvement and the theoretically countering force of consensus bias vary with public attention to disclosed ratings. The results enable managers to account for public attention to disclosure in predicting its performance and bias effects.

2.2 THEORY AND MOTIVATING LITERATURE

Economic theory offers reasons to expect positive performance effects of consumer-rating disclosure. Consumer ratings influence purchasing decisions in many industries, including health care (Chevalier and Mayzlin [2006], Dranove and Dafny [2008], Hannauer et al. [2014], Luca [2016]). If consumer-rating disclosure creates a more competitive market in which favorable ratings lead to revenue, the disclosure should enhance incentives to perform well as measured by the ratings. Jin and Leslie [2003] provide evidence of disclosure's performance effects and attribute the effects to revenue-based incentives.

Consumer-rating disclosure may further incentivize performance through forces of social comparison. By facilitating the comparing of oneself to peers, disclosure bolsters the self-image of individuals who perform well relative to others (Brown et al. [2012], Smith [2000]). Tafkov [2013] provides lab evidence that disclosure elicits performance improvement via self-image incentives.

However, consumer-rating disclosure may impede performance improvement by biasing raters toward the published consensus rating in forming their own ratings ("consensus bias").

Consensus bias would dull the sensitivity of ratings to effort. Reduced sensitivity of a performance measure diminishes the economically optimal amount of effort to exert toward improvement by that measure (Banker and Datar [1989]). By obscuring the ratings' responsiveness to changes in service practices, consensus bias would also inhibit physicians' ability to detect the success or failure of an attempt to improve service. Accounting literature suggests that this inhibition of trial-and-error learning would have negative performance consequences (Campbell et al. [2011]). Prior literature on disclosure's performance effects has not addressed a disclosure of subjective ratings to raters, nor explored the existence of potentially resultant consensus bias and its implications for performance (e.g., Bennear and Olmstead [2008], Jin and Leslie [2013], Lu [2012]).

Behavioral economic theories inform the likelihood of consumer-rating disclosure yielding consensus bias. The anchoring and adjusting heuristic from behavioral economics, if applicable to consumer ratings, would contribute to such an effect. Tversky and Kahneman [1975] established evidence of this heuristic, whereby exposure to an arbitrary number draws subsequent estimations and predictions toward itself. Many studies have shown similar results whereby an arbitrary number (e.g., one that an individual sees upon spinning a "wheel of fortune") sways estimation or prediction regarding a subject of limited familiarity to the individual (e.g., the number of nations in Africa) (Furnham and Boo [2011], Tversky and Kahneman [1975]). Consumer ratings, though, are distinct from the reference points in those studies. A published rating consensus is informative, rather than arbitrary, which may strengthen its salience as a subconscious reference point for anchoring. On the other hand, consumers are relatively familiar with the service being rated, and may thereby rely less on a subconscious anchor to arrive at a rating. Indeed, Muchnik et al. [2003] find evidence that arbitrarily assigning

a single negative rating to user-contributed online content does not affect the average favorability of subsequent ratings, although it is unclear whether this holds for representative consensus ratings. The theoretical tension and extant empirical evidence do not allow a strong prediction of the application of anchoring to consumer-rating disclosure.

Social herding may also contribute to a consensus-bias effect. Social herding theory explains how members of a peer group converge toward the same decision by observing and conforming to others' decisions (Baddeley [2010]). Keynes [1930] applied this theory to understand speculation-driven financial market events. He argued that individuals conform to others' decisions because they 1) feel uncertain about the decision, and 2) assume others are better informed. The influence of consumer ratings on purchasing decisions speaks to consumers' likelihood of assuming others are well-informed (Chevalier and Mayzlin [2006], Hanauer et al. [2014], Luca [2016]). However, a consumer may not assume that others are better informed than him or herself, especially if the consumer views the rating as regarding his or her personal experience. Models of consumer behavior, and mixed lab evidence, suggest that consumers may exaggerate a rating after seeing one that they disagree with (Eryarsoy and Piramuthu [2014]). As with anchoring, it is not ex-ante clear whether social herding will apply to consumer-rating disclosure.

This paper advances six hypotheses. Hypotheses 1-3 state the predicted tradeoff whereby physician-rating disclosure yields performance improvement despite creating a consensus bias. Hypotheses 1 and 2 regard performance improvement, and Hypothesis 3 regards consensus bias.

H1: Physician-rating disclosure positively affects ratings.

H2: Physician-rating disclosure positively affects objectively measured care quality.

H3: Physician-rating disclosure biases subsequent raters toward the published consensus.

Testing these hypotheses extends three streams of literature. First, it informs economics-based research that assesses the performance effects of disclosure. Although consumer-rating disclosure is quickly spreading in service industries, little to no research has addressed its performance effects. H1 and H2 predict a performance improvement effect of physician-rating disclosure. The testing of H3 will indicate whether that effect occurs despite a consensus-bias effect.[10]

Further, I test for performance effects on customer satisfaction as well as on objective quality measures. Customer satisfaction and product/service quality generally lead to better financial performance, and accounting research documents difficulty in incentivizing such nonfinancial performance measures through financial contracts.[11] For instance, supervisors manipulate evaluations to avoid political costs, and exhibit an aversion to basing compensation on nonfinancial measures (Bol et al. [2010], Ittner et al. [2003]). The current study assesses consumer-rating disclosure as an alternative means of incentivizing nonfinancial performance measure improvement.

A second stream of literature that this study extends, through the tests of H1 and H2, is that of health care disclosure research, which has so far found generally disappointing effects

[10] Positive time trends in the consumer ratings, shown in Table 3 Column 3, cause consensus bias to counter rating improvement in actuality and as estimated via difference-in-differences.

[11] Ittner and Larcker [1998] find evidence of customer satisfaction leading to financial performance and higher market valuations in a nonlinear manner, Nagar and Rajan [2005] find a strong relationship between customer satisfaction and future profits, and Banker et al. [2000] find positive effects of both quality and customer satisfaction on financial performance. Also, high levels of customer satisfaction that are publicly visible via awards and/or online ratings have been shown to drive financial performance and market returns (Balasubramanian et al. [2005], Chevalier and Mayzlin [2006]).

(Dranove et al. [2003], Epstein [2006], Shukla [2013], Ryan et al. [2012]). Coronary artery bypass graft (CABG) mortality rate disclosure, now mandated in 15 states, is the longest-standing and one of the most widely studied variants (Shukla [2013]). Some studies find declining mortality rates following disclosure (e.g., Hannan, Sarrazin, Doran, and Rosenthal [2013], Peterson, DeLong, Jollis, Muhlbaier, and Mark [1998]), but are unable to distinguish this from the results of contemporaneous initiatives (Shukla [2013]). Dranove et al. [2003] explain that observed declines in mortality rates may be due to changes in patient characteristics rather than in care quality, and show that disclosure leads physicians to select against unhealthy patients. A metastudy of mortality-rate disclosure finds generally inconclusive performance effects, and recent studies find little to no performance effects (Epstein [2006], Ryan et al. [2012], Shukla [2013]). Scholars also report adverse effects of disclosing hospital-level costs, including insurers' using the information to negotiate away discounts for competitors, and hospitals' manipulating nominal charges to obscure actual relative cost-efficiency (Christensen, Floyd, and Maffett [2014], Cutler and Dafny [2011]).

The practice of disclosing ratings of individual physicians is relatively recent. In 2012, UUHC became the first academic health care system to make such a disclosure online. Others including Stanford Health Care, Wake Forest Baptist Health, and Cleveland Clinic have followed. Studies suggest that a lack of patient and referring-provider attention to mortality-rate disclosure weakens its impact on competition and resulting performance improvement (Brown et al. [2012], Epstein [2006], Kolstad [2013]). Survey and field evidence indicates that physician ratings receive significantly more attention (Brown et al. [2012], Hanauer et al. [2014]). Further, a study of health insurance markets reported that, in selecting health plans, individuals respond to others' subjective ratings of care, but not to objective quality measures (Dafny and Dranove

[2008]). Physician-rating disclosure may, by virtue of its reported influence on patient choice, be more likely to result in performance improvement than the objective quality measure disclosures assessed in the referenced studies. I provide evidence of a performance effect.

By testing H3, regarding consensus bias, this study extends a third stream of literature, regarding the interpretation and application of consumer ratings. Ratings biased toward a published consensus rating would lag a move to a new service level in depicting that new service level. The ratings would, in part, reiterate prior ratings that regarded a different service level. The ratings would more fully depict the new service level as ratings from the period of updated service became a greater portion of the published consensus rating. Such a lag is relevant to research on the sequence of service improvement, online consumer ratings, and financial performance (Banker et al. [2000], Banker and Mashruwala [2009], Luca [2016]). Managers who account for the lag could more accurately anticipate the timeline for a service improvement to yield its full effect on consumer ratings. Researchers could similarly view online consumer ratings as lagged, rather than coincident, indicators of service improvement. This would help researchers to avoid measurement error in the timing of service improvement and financial performance and to thereby more accurately estimate their causal relationship (Banker and Mashruwala [2009]).

Evidence of consensus bias also informs attempts to mitigate moral hazard. A variety of managers and regulators incentivize performance by consumer ratings that are visible to subsequent raters (Flaherty [2014], Lyu et al. [2013], Ubel [2015]). By accounting for a lag in those ratings' updating, principals could more precisely measure effort and provide rewards accordingly. For instance, academic promotion review boards could interpret student-evaluation improvement after an assistant professor received a negative consensus evaluation that was made

visible to his or her subsequent students as requiring greater effort than if the consensus were kept private. If such effort went underestimated, the assistant professor would be incrementally subject to moral hazard, specifically, the tendency to avoid exerting unobservable effort (Banker and Datar [1989], Hölmstrom [1979]).

Finally, evidence of consensus bias is relevant to research on bias in employee evaluations. Extant research on employee evaluation bias addresses issues of leniency, centrality, race, and favoritism, but has not addressed a bias toward past consensus evaluations (Bol [2011], Moers [2005], Prendergast and Topel [1993]). Consensus bias from consumers could affect employee evaluations that incorporate consumer ratings that are visible to subsequent raters, as in the academic promotion example. Consensus bias may also affect employee evaluations through supervisors who view past consensuses and are biased toward them in evaluating an employee (Grote [2015]).

To extend this paper's contribution to performance-disclosure and consumer-rating research, I address how the performance-improvement and consensus-bias effects vary with recent public attention to consumer-rating disclosure. Public attention would, in theory, reinforce disclosure's performance effects by increasing the financial and self-image consequences of performing well by disclosed measures (Graham [2000], Parker and Nielsen [2011], Weil et al. [2006]). Hypotheses 4 and 5 address whether stronger rating improvement and objective-quality-measure improvement effects follow greater public attention to disclosed consumer ratings.

Consensus bias is also predictably stronger following recent public attention to consumer-rating disclosure. In particular, anchoring and social herding would be more likely to occur if a larger number of individuals viewed the published consensus and each viewed it

enough times to recall it. Hypothesis 6 addresses whether stronger consensus bias follows greater recent public attention to disclosed consumer ratings.

H4: The positive effect of physician-rating disclosure on ratings is greater under conditions of greater recent public attention to the disclosure.

H5: The positive effect of physician-rating disclosure on objectively measured care quality is greater under conditions of greater recent public attention to the disclosure.

H6: Physician-rating disclosure biases subsequent raters toward the published consensus to a greater degree under conditions of greater recent public attention to the disclosure.

Testing H4-H5 offers some of the first evidence of a correlation between recent public attention to disclosure and performance effects. A positive correlation would support theory that public attention strengthens performance incentives tied to disclosure (Parker and Nielsen [2011], Weil et al. [2006]). The test of H6, regarding consensus bias under greater public attention, makes two additional contributions. First, it identifies public attention as a contextual factor that warrants more significantly adjusting for consensus bias in interpreting and applying publicly visible consumer ratings. Second, it assesses a potential impediment to a positive correlation between public attention and subsequent performance improvement as measured by ratings. Specifically, public attention to a consensus rating may strengthen consensus bias and relatedly counter efforts to lift ratings above the published consensus.

2.3 SETTING

The field research site, UUHC, is an academic medical system comprised of four hospitals, eleven community clinics, and several specialty centers. The system receives over 1.4 million outpatient and over 30,000 inpatient visits annually, offering services ranging from primary care to the most advanced types of cancer treatment. UUHC is a leader in health care quality and safety, placing in the top 10 on the University Health Consortium's ranking of approximately 120 U.S. academic medical centers for six years in a row. It has recently prioritized patient satisfaction improvement, climbing from the bottom quartile to above the 80th percentile of peer hospitals by satisfaction measures from 2008 to 2014.

2.3.1 Physician ratings as a performance measure

Government, insurers, and health care systems increasingly use patient ratings of physicians and care as performance measures. Private insurers and Medicare have, over the past decade, incorporated the ratings into reimbursement calculations and public performance reports at the hospital level (Lyu et al. [2013]). UUHC and other health care systems have created departments tasked with improving patient experience as measured by physician ratings (Daniels and Lee [2014], Merlino and Raman [2013]).

2.3.2 Physician ratings at UUHC

UUHC had launched a patient satisfaction improvement initiative several years before disclosing physician ratings. Letters from unsatisfied patients were often addressed to UUHC's CEO, who remarked at the start of a patient satisfaction initiative in 2008, "If we are doing the right things, why are the patients so unhappy? I want it to feel better here" (Daniels and Miller [2014]). As part of the initiative, UUHC subscribed to Press Ganey Inc., the nation's largest patient satisfaction survey vendor. Press Ganey began automatically emailing a patient satisfaction survey following each UUHC patient visit, including during the entirety of the current study's

sample period. The survey asks patients to rate physicians on several criteria, listed in Appendix A.

As UUHC's patient satisfaction improvement efforts progressed, commercial online physician rating systems challenged the organization to consider disclosing its own physician ratings publicly. In December 2012, UUHC became the first academic health care system to make such a disclosure.

2.3.3 Publicly disclosing physician ratings

Physician profile web pages, already integrated within the UUHC official website, served as the venue for physician ratings disclosure.[12] The only criterion for a physician's inclusion in disclosure was that he or she had received 30 or more ratings in the 12 months preceding a rating posting. Physicians who met that criterion in December 2012 were the first to have their ratings posted on their online profile. In July 2013, any physicians who failed to meet the criterion in December 2012, but met it in July 2013, similarly had their ratings disclosed. Disclosed ratings were the physicians' 12 month average prior to the most recent posting that the physician met the survey count criterion for.

Along with quantitative ratings, UUHC posted all comments regarding the physician that did not identify the patient or contain slander or profanity. The effects of consumer-rating disclosure herein refer to a disclosure that included such comments, which are a common element of consumer-rating disclosures (e.g., Cleveland Clinic [2016], Columbia University [2016], Starwood [2016]). Example comments appear in Appendix B. Though comments are not used as dependent variables, the example comments illustrate the type of qualitative information disclosed along with quantitative ratings.

[12] This paper refers to UUHC employees who conduct patient-visits as "physicians," although some do not hold a doctoral degree. Alternative degrees fall in nursing, assistant, and specialist categories.

UUHC administrators reported heightened physician interest in ratings after the ratings were disclosed. For example, a hospital executive not involved in proposing or designing the disclosure initiative remarked that the disclosure "had a dramatic impact on the culture of the physicians and their engagement [in patient satisfaction]".

2.3.4 Generalizability

I am limited to data from one organization. The results may, though, be relatively generalizable for a few reasons. First, a number of health care systems, including Stanford Health Care and Cleveland Clinic, have disclosed physician ratings on physicians' official online profiles in the same manner as UUHC. The intervention is thus similar to that spreading in health care. The same format of emailing surveys to gather responses that are periodically posted online is also similar to rating disclosures in other industries, such as higher education (e.g., Columbia University [2016]). Second, the data span tens of thousands of patients, hundreds of physicians, and nearly four years. They also span a large patient and physician geographic area. UUHC consists of four hospitals, ten community clinics, and several specialty centers that serve a referral area encompassing five surrounding states and over 10% of the continental U.S. A reason that the results may be limited, even relative to other studies regarding individual institutions, is that the disclosure was made in a culture that supported and offered training resources for rating improvement. The performance effects may partly depend on these contextual factors.

2.4 DATA

The sample consists of proprietary data regarding patient visits, patient satisfaction survey responses, and web traffic to physicians' official online profiles. I restrict the sample to

physicians present with UUHC in both the period before the first instance of disclosure and after the last. That restriction: 1) affords including physician fixed-effects in difference-in-differences models, and 2) orients the models to test for changes in physician behavior rather than changes in physician group composition. The test of consensus bias further restricts the sample to physicians who had received ratings more than one year before disclosure. This allows constructing the consensus from which to measure bias toward during the year prior to disclosure. The consensus-bias results are robust to stipulating that the consensuses consist of various numbers of survey responses, as described in Section 5.3. Table 1 contains descriptive statistics of patient, visit, procedure, and physician characteristics. Appendix C contains variable definitions.

2.4.1 Patient-visit characteristics

Patient-visit data fields include gender, age (with ages above 89 treated as 90, in compliance with privacy standards), whether the insurance provider was Medicare or Medicaid, whether the patient speaks English, and whether the patient was visiting the physician for his or her first time. I incorporate patient age as indicator variables to account for nonlinearity in the relationship between age and the models' dependent variables. [13] In the analyses of ratings and bias, this is a set from psychology research that represents differences in emotion and cognition that would plausibly influence formation of a rating (Newman and Newman [2014]). In the analysis of objectively measured quality, this is a set outlined by the Centers for Medicare and Medicaid (CMS) to adjust quality measures for patient health risks (CMS [2016]).

Additional patient-visit characteristics are two measures of severity and complexity of the patient's condition. One is the Medicare reimbursement weighting associated with the visit,

[13] Surveys regarding visits of patients who are too young or otherwise incapable of responding are sent to the caretaker whose email is associated with the patient's medical record. The age of respondents who are caretakers of the patient is not available in the data, and so is represented indirectly by the patient's age (e.g., an infant age group would control for the typical age range of individuals who are caretakers of infants).

which reflects the severity of the case and the related complexity of treatment.[14] The second is the Charlson Comorbidity Index (CCI), a weighted score that represents the disease burden of the patient.[15] The CCI takes a value of one, two, three, or six, in general proportion to the likelihood of mortality within one year associated with the comorbid condition. CCI conditions range from ulcers to cancer. The conditions are recorded at the time of a procedure. Objectively measured quality analyses regard procedures, and are thus able to include CCI as measured at the given visit. Ratings may occur as part of a visit whether or not it involves a procedure, and thereby whether or not CCI is measured. For rating analyses, I thus include CCI as its value for the patient in UUHC visits during the six-month window centered at the rated visit. The results are robust to narrowing this window to three months or expanding it to one year.

2.4.2 Physician characteristics

The data identify a physician's gender, possession of an MD, age, number of years employed by UUHC, and whether he or she is a tenure-track physician. The tenure track is available through UUHC's affiliation with the University of Utah Medical School. Generalized difference-in-differences analysis of an intervention with multiple implementation dates utilizes individual (herein physician) fixed effects. Physician fixed effects subsume time-invariant physician characteristics. The noted data on these characteristics, though, allow comparing the physician groups included in and excluded from disclosure, as well as constructing matched samples of physicians for robustness tests.

2.4.3 Web traffic

[14] See Brown [2003] and Evans, Hwang, and Nagarajan [2001] for similar application of this measure.
[15] See Sundararajan, Henderson, Perry, Muggivan, Quan, and Ghali [2004] for a description of the index, and Dafny [2005], Chandra, Gruber, and McKnight [2010], Doyle Jr. [2011] for examples of its use.

TABLE 1: SAMPLE SELECTION AND DESCRIPTIVE STATISTICS

Panel A - Sample selection

Ratings	
Initial observations	178,334
Exclude physicians who exit the sample before or enter the sample after the first rating posting	(69,154)
Sample for ratings	109,150
Exclude physicians who enter the sample more recently than a year before the first rating posting	(9,446)
Restrict sample to one year before first rating posting and to before the second rating posting	(59,228)
Sample for absolute difference	40,476
Procedures	
Initial observations	48,839
Exclude physicians who exit the sample before or enter the sample after the first rating posting	(7,232)
Sample for quality deductions	41,607

Panel B - Ratings descriptive statistics

	Physician included in disclosure			Physician excluded from disclosure		
Unit of observation	N	Mean	SD	N	Mean	SD
Patient Visit						
Gender	106,171	0.60	0.48	2,979	0.50	0.50
Age	106,171	49.84	19.80	2,979	44.64	23.92
English speaking	106,171	0.98	0.13	2,979	0.98	0.11
Charges ($)	106,171	281.56	1,181.75	2,979	300.25	862.08
Severity/complexity	106,171	1.94	2.95	2,979	2.08	8.06
Commorbidity	106,171	0.01	0.14	2,979	0.01	0.14
Medicare or Medicaid	106,171	0.17	0.37	2,979	0.16	0.37
First visit	106,171	0.25	0.43	2,979	0.31	0.46
Physician-week's visit count	106,171	56.12	41.55	2,979	37.45	36.21
Physician						
Gender	295	0.36	0.48	99	0.37	0.48
MD	295	0.79	0.40	99	0.84	0.36
Age	295	44.61	10.03	99	43.50	9.21
Years with UUHC	295	9.34	5.05	99	7.85	4.11
Tenure track	295	0.31	0.46	99	0.36	0.48
Physician-website Month						
Web traffic	9,962	79.12	75.29	807	51.35	45.11
Satisfaction Survey						
Rating	106,171	4.70	0.52	2,979	4.74	0.48
Absolute difference	106,171	0.36	0.35	2,979	0.31	0.33

Physician ages are as of January 1, 2011, and patient ages are treated as 90 if above 89 in compliance with privacy standards.

TABLE 1: SAMPLE SELECTION AND DESCRIPTIVE STATISTICS (CONTINUED)

Panel C - Procedures descriptive statistics

	Physician included in disclosure			Physician excluded from disclosure		
Visit characteristic	N	Mean	SD	N	Mean	SD
Patient Visit						
Gender	36,827	0.54	0.49	4,780	0.55	0.49
Age	36,827	49.28	21.51	4,780	43.80	23.66
Charges ($)	36,827	44,194.52	101,574.10	4,780	44,957.01	139,024.80
Severity/complexity	36,827	2.12	3.40	4,780	2.04	3.38
Commorbidity	36,827	0.12	0.46	4,780	0.25	0.72
Medicare or Medicaid	36,827	0.04	0.21	4,780	0.09	0.29
Physician-week's visit count	36,827	36.10	25.69	4,780	13.71	12.41
Physician						
Gender	158	0.30	0.46	36	0.31	0.46
MD	158	0.94	0.22	36	0.86	0.35
Age	158	43.52	8.79	36	39.05	6.74
Years with UUHC	158	8.87	4.27	36	6.55	3.74
Tenure track	158	0.45	0.49	36	0.27	0.45
Physician-website Month						
Web traffic	3,970	88.79	84.54	712	41.47	39.19
Quality						
Quality deduction	36,827	0.02	0.15	4,780	0.02	0.16

Physician ages are as of January 1, 2011, and patient ages above 89 are treated as 90 in compliance with privacy standards.

Available data regarding web traffic are the number of page views of a physician's online profile each month during the sample period. A page view is an instance of a web browser loading the physician's online profile.

2.4.4 Physician ratings

Physician ratings are patient responses to automated emails sent by a third-party company, Press Ganey Inc. The survey content and distribution is not subject to the discretion of the physician conducting the visit. Patients answer the questions on a Likert scale of 1-5 with 1 indicating "very poor," and 5 indicating "very good". Rating components are listed in Appendix A. The average rating for an individual visit by all nine questions is the current study's dependent variable *rating*. The average of *rating* over the 12 months prior to a date of disclosure is made visible atop his or her profile upon rating disclosure.

Rating is generally very high, with an average in the sample for this study of 4.7 out of 5. This raises the possibility that institutional factors lead only satisfied patients to respond to surveys. However, this high average is representative of hospitals nationally. For UUHC's peer group of 120 academic hospitals that use Press Ganey surveys, the average is roughly 4.6 out of 5. The automated emailing system that directs surveys only to confirmed patients would also make it difficult for physicians to manipulate the respondent pool relative to traditional online reviews or to surveys that a physician distributes.

2.4.5 Procedures and quality

The objectively measured quality analysis applies to visits during which the physician performed a procedure, regardless of whether the patient subsequently responded to a survey regarding the visit. The proxy for quality is *quality deduction*, an indicator for whether the visit resulted in either a patient readmission to the emergency department within 30 days, a hospital acquired

condition, or both. A decrease in *quality deduction* corresponds to an increase in quality. This variable's component measures—30-day readmissions and hospital acquired conditions (conditions acquired along with treatment at the hospital)—are measured by health care regulators and widely studied as adverse and avoidable events that indicate quality failures.[16] UUHC provided these component measures dating back to one month after the start of the patient satisfaction dataset, affording nearly two years of quality observations before physician-rating disclosure. The procedure data that UUHC provided included all of the variables described in Sections 4.1 and 4.2, except whether the visit was a patient's first to the provider and whether the patient indicated that they speak English. Those two fields are gathered and made available by Press Ganey to UUHC only along with survey response data

2.5 ANALYSIS

This study utilizes the criterion determining whether a physician's ratings were posted online, as well as the timing of the postings, in its generalized-difference-in-differences tests. The single criterion was that a physician received 30 or more survey responses in the 12-month period prior to a rating posting. Rating postings occurred twice during this study's sample, once in December 2012, and again in July 2013. Posted ratings stayed constant on the physician's profile until a subsequent posting that the physician met the survey count criterion for inclusion in. The estimation employs generalized difference-in-differences, which has been used in a prior study regarding health care disclosure (Dranove et al. [2003], Duflo [2002]). In addition to physician and time fixed effects, the models also include extensive controls for patient mix, including multiple measures of patients' health risks.

[16] See Andel, Davidow, Hollander, and Moreno [2012], and Joynt, Orav, and Jha [2011] for descriptions and applications.

A key identifying assumption is that each dependent variable would trend parallel for physicians included in and those excluded from disclosure were it not for the disclosure. The analysis includes placebo tests and graphical illustrations of these trends prior to the start of disclosure at UUHC. A second key assumption is a lack of contemporaneous changes at the time of disclosure that would otherwise alter the parallel trends. Propensity-score matching of physicians allows testing for robustness to contemporaneous changes that would arise due to differences in demographics of physicians included in and excluded from disclosure. Additional tests show that the effects are robust after excluding, and are of comparable size among, physicians who first met the criterion number of returned surveys within one year prior to their inclusion in disclosure. This suggests that potential contemporaneous changes in dependent variable trends due to a physician's recently meeting the criterion for disclosure do not drive the results.

2.5.1 Rating improvement

Model 1, specified as follows, tests for the effect of physician-rating disclosure on subsequent ratings:

$$(1)\ \text{Rating}_{pytv} = \alpha + \delta\text{physician}_p + \lambda\text{year}_y + \omega\text{period}_t + \varsigma\text{controls}_v + \beta\text{disclosed}_{pt} + \varepsilon_{pytv},$$

where p indexes physicians, y indexes years, t indexes time periods segmented by disclosure events, v indexes individual patient visits, and disclosed is an indicator for the time period following which a physician's ratings were disclosed, if ever, on their online profile. The dimensions of the coefficient vectors are δ (1 × 394), λ (1× 4), ω (1 × 3), ς (1 × 15), and β (1 × 1). Physician fixed effects control for physician characteristics, including membership in or

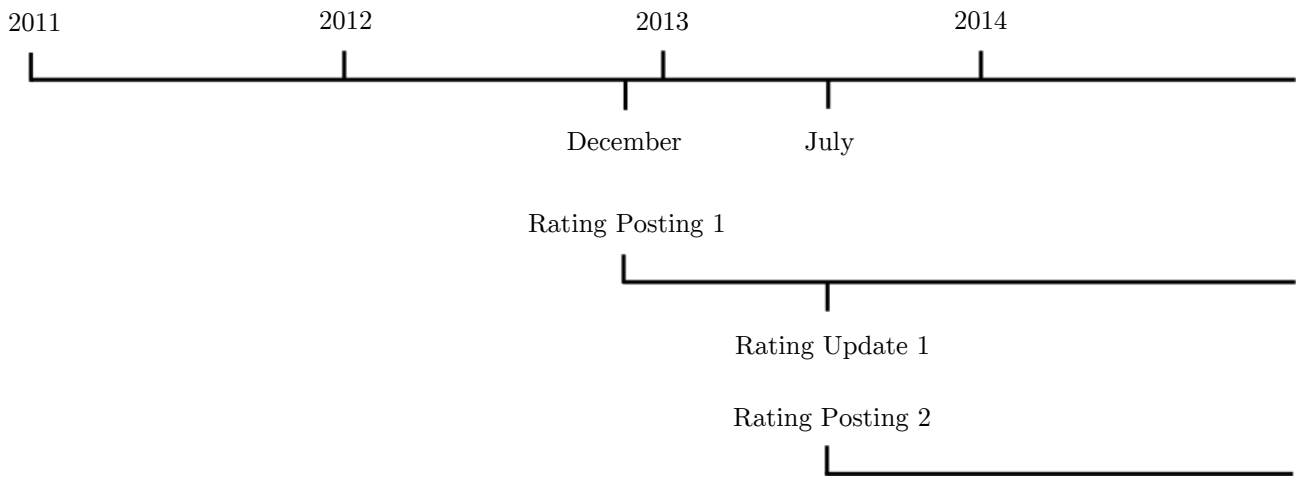
exclusion from the treated group. Time fixed effects control for time trends common to the entire sample, including differences between the before and after periods. The time variables are year and period, the latter of which controls for static variation prior to and after disclosure events occurring mid-calendar-year. B captures the difference-in-differences estimation of the effect of physician-rating disclosure on ratings. The unit of observation is the patient-visit. Data at the visit level allow controlling for observable patient and visit characteristics. These are the monetary charges for the visit, the severity/complexity of the case, the patient's comorbidities, whether the visit's insurer was Medicare or Medicaid, the patient's gender, a set of psychometric indicator variables for patient age, and an indicator variable for whether the visit was the patient's first to the physician. In estimating this model and all others in this study, standard errors are clustered at the physician level to correct for autocorrelation among multiple observations within physicians.[17]

Table 2 supports the parallel trends assumption. Column 1 displays results of estimating Model 1 using a placebo disclosure, timed one year prior to the date on which a physician's ratings were actually disclosed, and confined to the pre-disclosure period. The coefficient on *placebo disclosed* indicates no significant difference in *rating* trends in the pre-disclosure period. Figure 2 illustrates *rating* trends in the pre-disclosure period for physicians included in and excluded from disclosure.[18]

[17] The number of years is too few for reliable clustering by the standards of Petersen [2009]. Clustering standard errors by both years and physicians, though accordingly incorrect, slightly strengthens the results' statistical significance.

[18] The variance of scatter points in Figures 2-4 is greater for physicians excluded from relative to those included in disclosure. This is attributable to the smaller sample size for the former group. When I scale the variance to equate the number of observations in the former group with that of the latter, the variances do not statistically significantly differ. Widely cited difference-in-differences studies use treatment and control groups that differ in sample size by a large multiple – at least as large as 18 (Card and Krueger [1994], Dynarski [2002]). A sample size difference between treated and control group reduces effective sample size, but these studies do not mention it as a threat to difference-in-differences estimation. The results in Tables 3-5 hold, as shown in Table 7, after propensity-score

FIGURE 1: PHYSICIAN-RATING DISCLOSURE TIMELINE

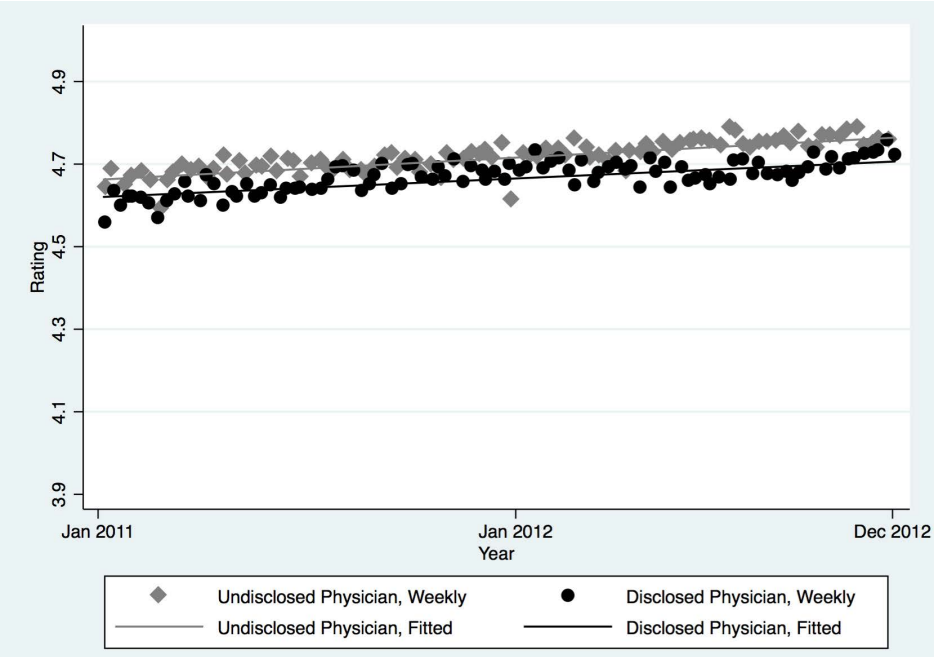


Rating Posting 1
Administrators posted, on the official webpage of each physician who had received at least 30 ratings during the prior 12 months, the physician's average rating for the nine questions in Appendix A during that period along with individual patient comments.

Rating Update 1
Administrators updated the ratings posted for physicians included in Rating Posting 1 and who had received at least 30 ratings during the prior 12 months to display those physicians' rating averages from that more recent 12-month period.

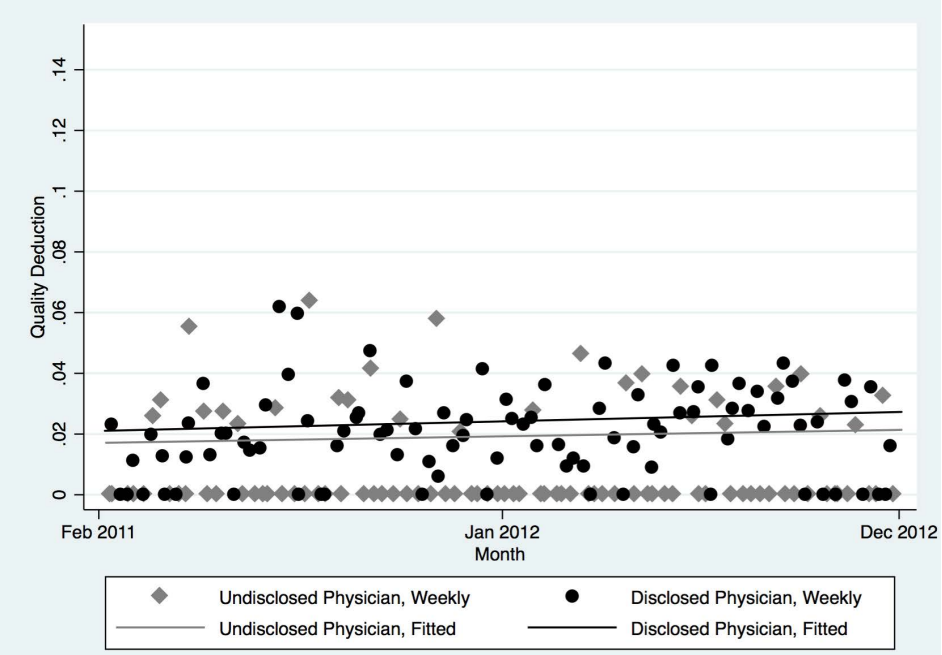
Rating Posting 2
Administrators posted, on the official webpage of each physician who had received at least 30 ratings during the prior 12 months, but who had not reached the threshold number of ratings for Rating Posting 1, the physician's average rating for the nine questions in Appendix A during that period along with individual patient comments.

FIGURE 2: PRE-DISCLOSURE RATING TRENDS BY PHYSICIAN'S EVENTUAL DISCLOSURE STATUS



This figure displays trends of *rating*, unadjusted for covariates, for physicians included in and excluded from ratings disclosure prior to the disclosure. Scatter points are agreggated at the week level. The scatter point variance for the group with fewer visit-level observations is reduced by the component of variance attributable to the smaller number of observations. Trend lines illustrate that *rating* was trending similarly for these groups of physicians prior to rating disclosure.

FIGURE 3: PRE-DISCLOSURE QUALITY DEDUCTION TRENDS BY PHYSICIAN'S EVENTUAL DISCLOSURE STATUS



This figure displays trends of *quality deduction*, unadjusted for covariates, for physicians included in and excluded from ratings disclosure prior to the disclosure. The variable's availability extends from February 2011. Scatter points are agreggated at the week level. The scatter point variance for the group with fewer visit-level observations is reduced by the component of variance attributable to the smaller number of observations. Trend lines illustrate that *quality deduction* was trending similarly for these groups of physicians prior to rating disclosure.

Table 2: Tests of Parallel Trends

	(1)	(2)	(3)
	Rating	Quality deduction	Absolute difference
Placebo disclosed	-0.020	0.003	-0.003
	[-0.83]	[0.67]	[-0.19]
Gender	-0.018**	0.001	0.007
	[-2.24]	[0.56]	[1.58]
Charges	0.000	-0.000*	0.000
	[0.78]	[-1.71]	[1.35]
Severity/complexity	0.001	0.001	-0.000
	[1.19]	[1.62]	[-0.50]
Commorbidity	0.037***	-0.002	-0.033***
	[3.05]	[-1.59]	[-4.46]
Medicare or Medicaid	0.012	0.002	0.005
	[1.55]	[0.49]	[0.92]
Physician-week's visit count	0.000	-0.000	-0.000
	[0.22]	[-0.41]	[-0.59]
First visit	-0.041***		0.025***
	[-5.42]		[4.00]
English speaking	0.087***		-0.067***
	[4.73]		[-3.52]
Contemporary std. dev.			0.315***
			[11.22]
Consensus count			0.045
			[0.49]
Rating trend			0.059***
			[3.38]
Age dummies°	Yes	Yes	Yes
Year dummy			
2012	0.014	0.008	-0.006
	[1.09]	[0.26]	[-0.70]
Period dummies			
2	0.039	-0.004	0.015
	[1.36]	[-0.76]	[0.76]
3	0.049	-0.000	
	[1.66]	[-0.11]	
Physician dummies	Yes	Yes	Yes

This table presents effect estimates of a placebo physician rating disclosure to assess, in the pre-disclosure period, paralell trends in ratings, quality deductions, and the absolute difference of ratings from prior consensus ratings. Placebo disclosure is timed one year before a physician's actual rating disclosure for the results in Columns 1-2. Data for constructing absolute difference is limited to one year before the absolute disclosure. Given that constraint, the placebo test for Column 3 times its placebo disclosure seven months prior to actual disclosure, which provides a post-treatment period for analysis of the placebo disclosure as long as that available for the actual disclosure, and provides a six month pre-treatment period for the placebo disclosure analysis. °Age dummies for Column 1 and 3 are the psychometric set in Newman and Newman [2014], and for Column 4 are the health-risk set delineated by CMS [2016], with one group omitted in each case. Standard errors are clustered at the physician level. *,**,*** denote significance at the .1, .05, and .01 levels respectively. Rating $N = 65{,}286$, Quality deduction $N = 27{,}456$, Absolute difference $N = 32{,}413$

Table 3 displays the results of estimating Model 1. The coefficient on *disclosed* in Columns 1 and 2 shows an estimated positive and statistically significant effect of the disclosure on improvement by patient satisfaction ratings. Column 3 shows positive time trends in *rating*. These trends are key to estimating improvement in *rating* that occurs in spite of, rather than due to, consensus bias. Consensus bias would draw *rating* toward its prior published value and thereby counter *rating*'s improvement in actuality and in difference-in-differences estimation with non-negative trends. The results from Columns 1 and 2 support H1 – that disclosing physician ratings positively affects ratings. The estimated effect is also economically significant in that it raises a physician's rank by the ratings among the national University Health Consortium peer group by 17 percentile points on average.[19] UUHC and other health care systems report these ranks to individual physicians privately and sometimes publicly in aggregate (Daniels and Miller [2014], Merlino and Raman [2013]).

2.5.2 Quality improvement

Model 2, specified as follows, measures the effect of physician-rating disclosure on the occurrence of objectively measured quality deductions including readmissions and hospital-acquired conditions:

$$(2)\ \text{Quality deduction}_{pytv} = \alpha + \delta\text{physician}_{p} + \lambda\text{year}_{y} + \omega\text{period}_{t} + \varsigma\text{controls}_{v} + \beta\text{disclosed}_{pt} + \varepsilon_{pytv}.$$

The model's subscripts and the right-hand variables, other than those contained in the controls vector, are the same as those in Model 1. The coefficient vector dimensions are δ (1 ×

matching reduces the multiple by which the sample sizes differ for physicians excluded from and included in disclosure to less than 13.

[19] The University Health Consortium percentile point conversions were produced using UUHC's internal data provided by Press Ganey that maps physician ratings to the consortium's peer group distribution.

Table 3: Effect of Physician Rating Disclosure on Rating

	(1)	(2)	(3)
		Rating	
Disclosed	0.032**	0.034***	
	[2.26]	[2.57]	
Gender		-0.016***	
		[-3.73]	
Charges		0.000	
		[0.91]	
Severity/complexity		0.002***	
		[3.08]	
Commorbidity		0.018*	
		[1.69]	
Medicare or Medicaid		0.014***	
		[2.70]	
Physician-week's visit count		-0.000	
		[-0.59]	
First visit		-0.048***	
		[-8.17]	
English speaking		0.089***	
		[5.41]	
Age dummies			
11-17		0.011	
		[0.54]	
18-24		-0.113***	
		[-4.49]	
25-34		-0.072***	
		[-3.23]	
35-59		0.012	
		[0.58]	
59-74		0.090	
		[4.12]	
+74		0.082	
		[3.66]	
Year dummies			
2012	0.043***	0.037***	0.044***
	[7.23]	[6.30]	[6.31]
2013	0.052***	0.046***	0.088***
	[3.45]	[3.12]	[11.05]
2014	0.065***	0.060***	0.101***
	[4.16]	[3.90]	[11.87]
Period dummies			
2	0.003	-0.003	
	[0.19]	[-0.20]	
3	0.006	-0.003	
	[0.29]	[-0.17]	
Physician dummies	Yes	Yes	No

This table presents effect estimates of physician rating disclosure on physician ratings. Columns 1-2 vary the controls included, and Column 3 presents isolated time trends. Standard errors are clustered at the physician level. Age dummies for Columns 1 and 3 are the psychometric set in Newman and Newman [2014]. *,**,*** denote significance at the .1, .05, and .01 levels respectively. $N = 109{,}050$

Table 3 displays the results of estimating Model 1. The coefficient on *disclosed* in Columns 1 and 2 shows an estimated positive and statistically significant effect of the disclosure on improvement by patient satisfaction ratings. Column 3 shows positive time trends in *rating*. These trends are key to estimating improvement in *rating* that occurs in spite of, rather than due to, consensus bias. Consensus bias would draw *rating* toward its prior published value and thereby counter *rating*'s improvement in actuality and in difference-in-differences estimation with non-negative trends. The results from Columns 1 and 2 support H1 – that disclosing physician ratings positively affects ratings. The estimated effect is also economically significant in that it raises a physician's rank by the ratings among the national University Health Consortium peer group by 17 percentile points on average.[19] UUHC and other health care systems report these ranks to individual physicians privately and sometimes publicly in aggregate (Daniels and Miller [2014], Merlino and Raman [2013]).

2.5.2 Quality improvement

Model 2, specified as follows, measures the effect of physician-rating disclosure on the occurrence of objectively measured quality deductions including readmissions and hospital-acquired conditions:

$$(2)\ \text{Quality deduction}_{pytv} = \alpha + \delta\text{physician}_{p} + \lambda\text{year}_{y} + \omega\text{period}_{t} + \varsigma\text{controls}_{v} + \beta\text{disclosed}_{pt} + \varepsilon_{pytv}.$$

The model's subscripts and the right-hand variables, other than those contained in the controls vector, are the same as those in Model 1. The coefficient vector dimensions are δ (1 ×

matching reduces the multiple by which the sample sizes differ for physicians excluded from and included in disclosure to less than 13.

[19] The University Health Consortium percentile point conversions were produced using UUHC's internal data provided by Press Ganey that maps physician ratings to the consortium's peer group distribution.

194), λ (1× 4), ω (1 × 3), ς (1 × 20), and β (1 × 1). The control vector comprises charges, the severity/complexity of the case, the patient's CCI comorbidity score, whether the visit's insurer was Medicare or Medicaid, the patient's gender, and the indicators for age as outlined by CMS for health risk adjustment.

The results in Table 2 support the parallel trends assumption for *quality deduction*. Column 2 displays results of estimating Model 2 using a placebo disclosure, timed one year prior to the date on which a physician's ratings were actually disclosed, and confined to the pre-disclosure period. The coefficient on *placebo disclosed* indicates no significant difference in *quality deduction* trends in the pre-disclosure period. Figure 3 illustrates the *quality* trends in the pre-disclosure period for physicians included in and excluded from disclosure.

Table 4 displays the results of the test of Model 2. I use ordinary least squares regression for estimating effects on a binary outcome variable, in line with prior health economics research (Dranove [2003]). This approach prohibits interpreting the results as point estimates, but avoids Type I error prone to result from applying logit or probit regression to difference-in-differences estimation (Blundell and Dias [2009]). The coefficient on *disclosed* in Columns 1-2 shows an estimated negative and statistically significant effect of the disclosure on quality deductions. The results support H2 – that disclosing physician ratings positively affects the quality of care that the physician provides.

2.5.3 Consensus bias

Model 3, specified as follows, measures the effect of physician-rating disclosure on consensus bias in subsequent ratings:

TABLE 4: EFFECT OF PHYSICIAN RATING DISCLOSURE ON QUALITY DEDUCTION

	(1)	(2)
	Quality deduction	
Disclosed	-0.010***	-0.010***
	[-2.61]	[-2.62]
Gender		0.004***
		[2.80]
Charges		-0.000
		[-1.11]
Severity/complexity		0.000
		[0.56]
Commorbidity		-0.002*
		[-1.74]
Medicare or Medicaid		0.006
		[1.50]
Physician-week's visit count		0.000
		[0.12]
Age dummies°		
1-9		0.002
		[0.18]
9-20		0.034**
		[2.49]
21-29		0.021*
		[1.77]
30-39		0.021*
		[1.74]
40-49		0.021*
		[1.72]
50-59		0.014
		[1.16]
+59		0.009
		[0.76]
Year dummies		
2012	0.002	0.002
	[1.57]	[1.63]
2013	0.002	0.002
	[0.51]	[0.58]
2014	-0.003	-0.003
	[-0.67]	[-0.55]
Period dummies		
2	0.007	0.007
	[1.41]	[1.42]
3	0.016**	0.016**
	[2.56]	[2.56]
Physician dummies	Yes	Yes

This table presents effect estimates of physician rating disclosure on quality deduction. Columns 1-2 vary the controls included. °Age dummies are the 14 categories delineated by CMS [2016] as capturing age-dependent health risk, with the first ommitted, and the subsequent 12 paired with their adjacent category for concise display. Decoupling the paired categories and including them as separate dummies in the regression does not alter the level of significance of the estimated effect of disclosure. Standard errors are clustered at the physician level. *,**,*** denote significance at the .1, .05, and .01 levels respectively. $N = 41{,}607$

(3) Absolute difference$_{pytv}$ = α + δphysician$_p$ + λyear$_y$ + ωperiod$_t$ + ςcontrols$_v$ + βdisclosed$_{pt}$ + ε$_{pytv}$.

The dependent variable, *absolute difference*, is the absolute distance of *rating* for the visit from the physician's consensus rating as calculated for rating disclosure. A reduction in *absolute difference*, controlling for a change in the overall standard deviation of a physician's ratings, is the proxy for consensus bias. Though UUHC did not calculate the measure prior to disclosure, the necessary information is calculable. For the year immediately preceding disclosure, I construct this measure as the absolute difference of *rating* from the physician's 12-month consensus as would have been calculated for disclosure one year before the disclosure's actual start. The variable's measurement, which requires establishing a 12-month consensus from which to measure absolute distance, begins as early as UUHC's data allows – in December 2011. In this test, the sample is also truncated at the time of the second disclosure, in July 2013, when physicians were added and initial ratings were updated. Beyond that point, the combination of consensus ratings for a physician that a patient rater may have been exposed to makes the appropriate standard for measuring bias unclear. The model's subscripts and the right-hand variables other than controls are the same as in Model 1. The coefficient vector dimensions are δ (1 × 273), λ (1× 1), ω (1 × 1), ς (1 × 18), and β (1 × 1). The control vector includes all controls used in Model 1 and three additional controls.

The first of these additional controls teases out the effect of a change in the overall standard deviation of *rating* for a given physician following disclosure. This control, *contemporary standard deviation* is the standard deviation of *rating* for a given physician in the period relative to the December 2012 posting—the only instance of disclosure used to test Model

3—in which the rating occurred. This control captures a physician becoming more or less consistent in the level of service he or she provides, or any other reasons for a decrease in the variance of his or her ratings not attributable to other observable covariates.

The second additional control, *consensus count*, teases out the effect of a difference in the rating sample size for a physician's consensus rating as calculated at December 2011 relative to December 2012. *Absolute difference* is measured relative to the former in the pre-disclosure period and relative to the latter in the post-disclosure period. *Consensus count* is the inverse square root of the sample size of ratings that constitute the physician's consensus rating from which *absolute difference* is measured. The sample mean of an i.i.d. random variable converges to the population mean is $1/\sqrt{n}$ (Vives p. 386 [2010]). *Consensus count* uses that transformation to capture an effect on *absolute difference* of a physician's December 2012 consensus rating being closer to or farther from his or her population mean rating than his or her December 2011 consensus rating was.

The third additional control, *rating trend*, teases out the effect of a physician's mean change in *rating* from the variance of *rating* around the published score. This control is the physician-specific trend for *rating* in the period relative to the December 2012 posting in which an observed rating occurred. Given that disclosure yields a positive effect on *rating* and that *rating* time trends are positive, excluding this control would cause *absolute difference* to relatedly grow more for physicians included in disclosure, and thereby understate consensus bias as measured by a decline in *absolute difference*.

Table 2 Column 3 displays results of a test of parallel trends of the dependent variable *absolute difference*. The start of the variable's measurement one year prior to disclosure does not allow estimating the effects of a placebo disclosure occurring that far in advance of actual

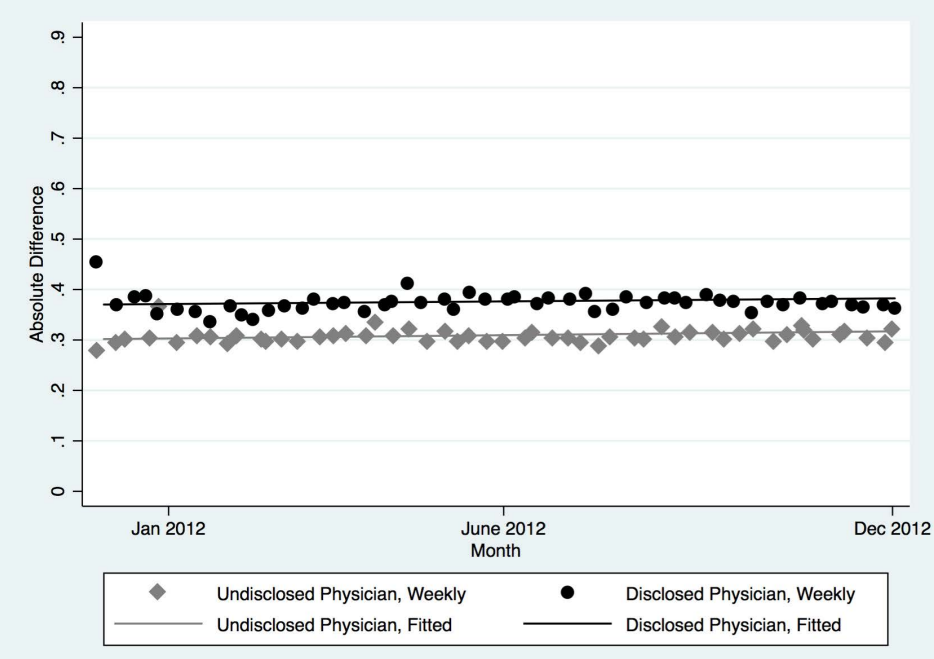
disclosure, given the necessity of a pre-period for the placebo disclosure. The placebo test uses instead a placebo disclosure date seven months prior to the date on which a physician's ratings were actually disclosed, which affords as long a period for assessment after placebo disclosure as is available for assessment after actual disclosure. The placebo test's sample is, as with the other placebo tests, confined to the pre-disclosure period. I update the consensus rating to the 12-month average for each physician at the time of the placebo disclosure, as would have occurred if the disclosure were actual. The coefficient on *placebo disclosed* indicates no significant difference in *absolute difference* trends. Attempting to depict parallel trends in attraction to the consensus score by displaying *absolute difference*, unadjusted for covariates, should be done with caution given that the control for *contemporary standard deviation* loads heavily. Figure 4, though, illustrates the trends for *absolute difference* as calculated for the test of model 3 during the pre-disclosure period.

Table 5 displays the results of the test of Model 3. The coefficient on *disclosed* in Columns 1-3 show an estimated negative and statistically significant effect of disclosure on *absolute difference*.[20] Columns 2 and 3 show stronger effects, as predicted, after controlling for physician-specific rating trends. The results support H3 – that disclosing physician ratings biases raters toward the published consensus. The estimated effect is economically significant in that the bias results in ratings that are an average of 24 percentile points closer to the physician's published consensus rating in the University Health Consortium peer group distribution.[21] Figure 5 illustrates the effect of eventual inclusion in disclosure on a physician's rating level and spread.

[20] The result is statistically significant at least the .05 level under various thresholds for a physician's inclusion in the sample based on the number of surveys comprising their consensus ratings, including >5, >10, >15, >20, and >25 surveys.

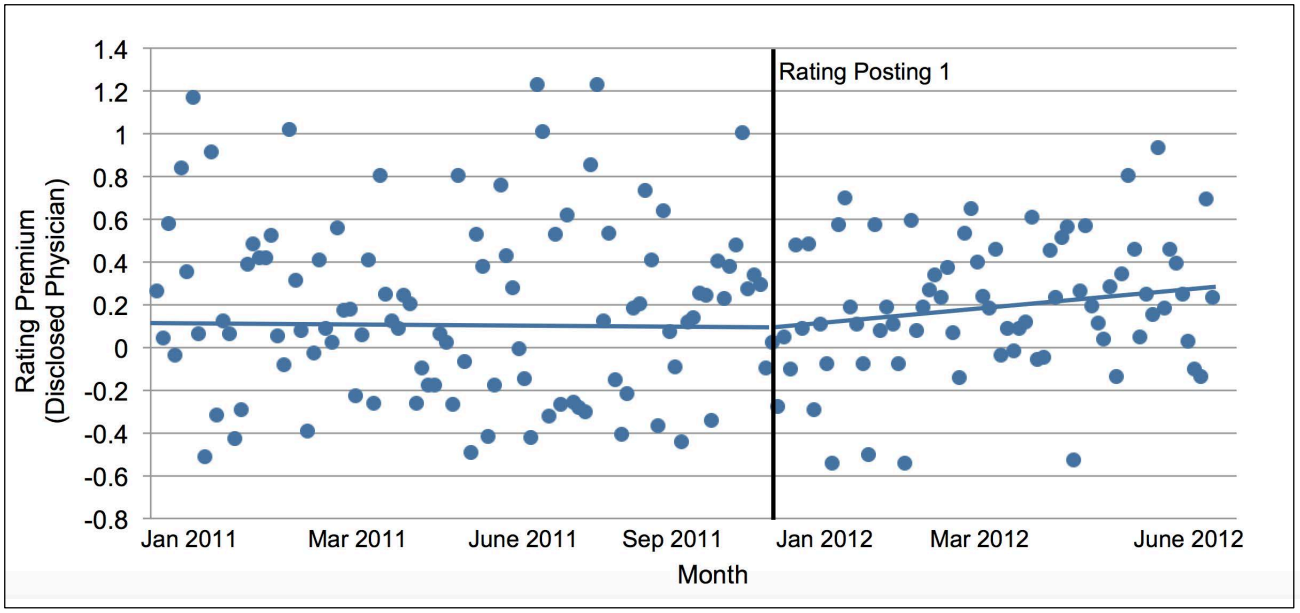
[21] The University Health Consortium percentile point conversions were produced using UUHC's internal data provided by Press Ganey that maps physician ratings to the consortium's peer group distribution.

FIGURE 4: PRE-DISCLOSURE ABSOLUTE DIFFERENCE TRENDS BY PHYSICIAN'S EVENTUAL DISCLOSURE STATUS



This figure displays trends of *absolute difference*, unadjusted for covariates, for physicians included in and excluded from ratings disclosure prior to the disclosure. Scatter points are aggregated at the week level. The scatter point variance for the group with fewer visit-level observations is reduced by the component of variance attributable to the smaller number of observations. *Absolute difference* is measured from the physician's 12-month consensus rating calculated at December 2011, the earliest point of that measure's availability and one year prior to ratings disclosure. Trend lines illustrate that *absolute difference* was trending similarly for these groups of physicians prior to ratings disclosure.

FIGURE 5: RATING PREMIUM FOR PHYISICIANS EVENTUALLY INCLUDED IN DISCLOSURE



This figure displays the effect of eventual inclusion in disclosure on a physician's *rating* level and spread. Trend lines show the average rating premium over time for a physician's belonging to the group eventually included in disclosure, after controlling for all covariates in Model 1. The lines are segmented to show a change in trend at Rating Posting 1. The variance of the scatter points in each quarter is the average variance of all physician ratings in the given quarter plus the estimated effect of consensus bias in that quarter on the variance of those ratings.

TABLE 5: EFFECT OF PHYSICIAN RATING DISCLOSURE ON ABSOLUTE DIFFERENCE

	(1)	(2)	(3)
	Absolute difference		
Disclosed	-0.040***	-0.046***	-0.047***
	[-2.92]	[-3.46]	[-3.61]
Gender			0.006
			[1.44]
Charges			0.000**
			[2.07]
Severity/complexity			-0.001**
			[-2.32]
Commorbidity			-0.017**
			[-2.05]
Medicare or Medicaid			0.003
			[0.66]
First visit			0.024***
			[4.87]
Physician-week's visit count			0.000
			[1.39]
English speaking			-0.074***
			[-4.29]
Contemporary std. dev.	0.374***	0.377***	0.374***
	[13.89]	[14.23]	[13.91]
Consensus count	0.137	0.130	0.129
	[1.11]	[1.13]	[1.09]
Rating trend		0.050***	0.052***
		[2.85]	[2.95]
Age dummies			
11-17			0.006
			[0.33]
18-24			0.076***
			[3.72]
25-34			0.049***
			[2.85]
35-59			0.016
			[1.07]
59-74			-0.016
			[-1.05]
+74			-0.024
			[-1.53]
Year dummies			
2012	-0.007	-0.006	-0.003
	[-0.68]	[-0.67]	[-0.38]
2013	-0.013	-0.013	-0.012
	[-1.00]	[-1.02]	[-0.93]
Period dummy			
2	0.035**	0.051***	0.056***
	[2.22]	[3.11]	[3.44]
Physician dummies	Yes	Yes	Yes

This table presents effect estimates of physician rating disclosure on the absolute distance of ratings from prior consensus ratings as were calculated for online disclosure. Columns 1-3 vary the controls included. Age dummies are the psychometric set in Newman and Newman [2014]. Standard errors are clustered at the physician level. *,**,*** denote significance at the .1, .05, and .01 levels respectively. $N = 39{,}159$

Relative to physicians excluded from disclosure, the ratings for physicians in this group begin trending upward and become more tightly distributed after the first rating posting.

2.5.4 Robustness tests

Table 7 demonstrates the robustness of the effects in sections 5.1-5.3 to estimation after matching physicians included in and excluded from disclosure by observable physician covariates: *age*, *MD*, *gender*, *years with UUHC*, and *tenure track*. The analyses so far have controlled for underlying physician differences in two regards. First, physician fixed effects control for the time-invariant effects of both observable and unobservable physician characteristics. Second, the placebo tests establish that time-variant effects of those characteristics on the dependent variables do not impede generally parallel trends.

Propensity-score matching on the noted observable physician covariates further evidences the results' robustness to observable physician characteristics by showing that the effects persist among non-significant differences in these characteristics between treated and control physicians. The sample matching utilizes a probit model that estimates propensity scores based on the observable characteristics. Subsequently, each member of the smaller group (physicians excluded from disclosure) is matched, without replacement, to the member of the larger group (physicians included in disclosure) with the most similar propensity score that has not yet been matched. For the test of *absolute difference*, the group of physicians excluded from disclosure are all those who did not meet the criterion for the December 2012 posting. Table 6 shows the covariate balance after matching. Table 7 shows that the results for *rating*, *quality deduction*, and *absolute difference* hold using these matched, and much smaller, samples. Each result is statistically significant at the .05 level or lower.

As an additional robustness test to help rule out contemporaneous changes in dependent variable trends, Table 8 shows effect estimates partitioned by whether the physician met the criterion for disclosure within the year prior to his or her inclusion in disclosure. The coefficients on *disclosure* in Columns 1, 3, and 5, which exclude physicians who met the criterion within a year prior to disclosure, are statistically significant at the .05 level or lower. The estimated effects among physicians who met the criterion within a year prior to disclosure, shown in Columns 2, 4, and 6, are of similar magnitude or lower and in no case more statistically significant. This test suggests that the effects of disclosure on *rating, quality deduction*, and *absolute difference* are not explained by a contemporaneous change in dependent variable trends upon a physician meeting the criterion number of surveys for inclusion in disclosure.

2.5.5 Public attention's moderation of disclosure's effects

Table 9 shows the disclosure effect estimates from Models 1-3 partitioned by recent public attention.[22] The proxy for recent public attention is *web traffic* – the number of page views of a physician's profile in the calendar month prior to a given patient visit.[23] Column 1 reports the partitioned effects on *rating*. χ^2 tests of differences in coefficients indicate a difference in the effects on rating, but in the opposite direction as predicted in H4; the disclosure's positive effect on *rating* is strongest at lower levels of recent public attention. This runs counter to theory from disclosure literature that public attention should strengthen disclosure's performance effects (Graham [2000], Parker and Nielsen [2011], Weil et al. [2006]). It may be partly explained by public attention to a disclosed rating reinforcing consensus bias and relatedly impeding rating

[22] Physician fixed effects subsume a physician's 12-month consensus rating at the time of disclosure, which is potentially correlated with web traffic as well as well as the disclosure's effects. Including this control accordingly does not affect the results of χ^2 tests for comparing effect size between partitions.

[23] The statistically significant results of χ^2 tests for comparing effect size between partitions are robust to using 3-calendar-month and 6-calendar-month-lagged web traffic in the partitioning. The level of significance in some cases differs, depending on the on the length of lagged used, but remains below 0.1.

TABLE 6: PROPENSITY-SCORE-MATCHED PHYSICIAN SAMPLE DESCRIPTIVE STATISTICS

	Physician included in disclosure			Physician excluded from disclosure		
Physician Characteristic	N	Mean	SD	N	Mean	SD
Rating Analysis						
Gender	99	0.33	0.47	99	0.37	0.48
MD	99	0.84	0.36	99	0.84	0.36
Age	99	42.62	9.41	99	43.50	9.21
Years with UUHC	99	7.77	4.33	99	7.85	4.11
Tenure	99	0.38	0.48	99	0.36	0.48
Quality Deduction Analysis						
Gender	36	0.27	0.45	36	0.33	0.47
MD	36	0.88	0.31	36	0.86	0.35
Age	36	38.11	7.26	36	39.05	6.74
Years with UUHC	36	6.35	3.47	36	6.55	3.74
Tenure	36	0.33	0.47	36	0.27	0.45
Absolute Difference Analysis						
Gender	115	0.36	0.48	115	0.36	0.48
MD	115	0.81	0.38	115	0.82	0.38
Age	115	43.52	10.27	115	43.33	8.98
Years with UUHC	115	7.65	3.74	115	7.61	4.13
Tenure	115	0.32	0.46	115	0.33	0.47

This table shows descriptive statistics of physicians comprising the samples used for testing the robustness of physician-rating disclosure effect estimates to matching physicians on gender, posession of an MD, age, number of years employed by UUHC, and status as a tenure track employee. The rating and quality deduction samples were produced using one-to-one propensity score matching, applied to match either of the two rating postings with those excluded from disclosure. The sample for absolute difference was produced by using the same matching procedure, applied to match physicians included in the first upload to those included in a later upload or excluded from disclosure. Physician ages are as of January 1, 2011. The covariates in the resulting samples do not exhibit statistically significant differences between matched groups, and all are within a 0.05 propensity-score caliper, as used in prior research pairing propensity-score matching with difference-in-differences estimation (e.g., Sandino and Murphy, 2010).

TABLE 7: ROBUSTNESS TEST USING PROPENSITY-SCORE-MATCHED PHYSICIAN SAMPLES

	(1)	(2)	(3)
	Rating	Quality deduction	Absolute difference
Disclosed	0.033**	-0.012**	-0.044***
	[2.36]	[-2.24]	[-2.97]
Gender	-0.013**	0.001	0.008
	[-2.15]	[0.76]	[1.21]
Charges	0.000	0.000	0.000**
	[0.42]	[0.41]	[2.46]
Severity/complexity	0.001	-0.000***	-0.009
	[1.80]	[-3.19]	[-1.19]
Commorbidity	0.022	-0.003	-0.021
	[1.17]	[-1.28]	[-1.15]
Medicare or Medicaid	0.017**	0.005	0.003
	[2.56]	[0.77]	[0.52]
Physician-week's visit count	-0.000	-0.000	0.000*
	[-0.11]	[-0.14]	[1.96]
First visit	-0.038***		0.014***
	[-4.29]		[1.98]
English speaking	0.069**		-0.036**
	[2.39]		[-2.43]
Contemporary std. dev.			0.373***
			[8.90]
Consensus count			0.164
			[1.03]
Rating trend			0.056*
			[1.97]
Age dummies°	Yes	Yes	Yes
Year dummies			
2012	0.033***	-0.000	-0.001
	[3.46]	[-0.23]	[-0.07]
2013	0.051**	-0.000	-0.003
	[2.27]	[-0.03]	[0.18]
2014	0.077***	-0.007	
	[3.23]	[-0.74]	
Period dummies			
2	-0.008	0.008	0.042**
	[-0.39]	[1.03]	[2.18]
3	-0.009	0.017	
	[-0.41]	[1.61]	
Physician dummies	Yes	Yes	Yes

This table presents effect estimates of physician rating disclosure on ratings, quality deductions, and the absolute difference of ratings from prior consensus ratings, using propensity-score-matched samples of physicians included in and excluded from the assessed disclosure events. °Age dummies for columns 1-2 and 5-6 are the psychometric set in Newman and Newman [2014], and for columns 3-4 are the set delineated by CMS (2016) for use in risk adjustment, with one group omitted in each case. Standard errors are clustered at the physician level. *,**,*** denote significance at the .1, .05, and .01 levels respectively. Rating $N = 37{,}741$, Quality deduction $N = 9{,}861$, Absolute difference $N = 15{,}501$

TABLE 8: ROBUSTNESS TESTS FOR PHYSICIAN'S RECENCY OF MEETING CRITERION FOR DISCLOSURE

	(1)	(2)	(3)	(4)	(5)	(6)
	Rating		Quality deduction		Absolute difference	
	>= 1 Year	< 1 Year	>= 1 Year	< 1 Year	>= 1 Year	< 1 Year
Disclosed	0.037***	0.024	-0.010**	-0.013**	-0.049***	-0.048***
	[2.64]	[1.09]	[-2.41]	[-2.33]	[-3.73]	[-2.30]
Gender	-0.017***	-0.003	0.003*	0.005**	0.008*	-0.006
	[-3.67]	[-0.22]	[1.82]	[2.03]	[1.87]	[-0.49]
Charges	0.000	0.000***	-0.000	-0.000	0.000**	-0.000
	[0.35]	[3.58]	[-0.43]	[-0.42]	[2.11]	[-0.01]
Severity/complexity	0.002***	0.000	0.001	-0.001***	-0.001**	0.000
	[2.85]	[0.52]	[1.04]	[-2.78]	[-2.20]	[0.04]
Commorbidity	0.021*	-0.001	-0.003*	-0.003	-0.022***	0.025
	[1.82]	[-0.02]	[-1.76]	[-1.34]	[-2.66]	[0.60]
Medicare or Medicaid	0.014**	0.024**	0.007	0.001	0.004	-0.012
	[2.41]	[2.05]	[1.44]	[0.22]	[0.80]	[-0.89]
Physician-week's visit count	-0.000	-0.000	0.000	-0.000	0.000	0.006*
	[-0.49]	[-0.43]	[0.27]	[-0.33]	[1.13]	[1.75]
First visit	-0.047***	-0.048***			0.023***	0.033**
	[-7.60]	[-3.05]			[4.55]	[2.36]
English speaking	0.089***	0.090*			-0.074***	-0.076
	[5.15]	[1.80]			[-4.21]	[-1.26]
Contemporary std. dev.					0.363***	0.396***
					[12.13]	[8.33]
Consensus count					0.432**	-0.020
					[2.53]	[-0.13]
Rating trend					0.044**	0.061*
					[2.25]	[1.66]
Age dummies°	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies						
2012	0.037***	0.053***	0.003	0.002	-0.000	-0.050
	[5.83]	[3.70]	[1.62]	[0.50]	[-0.04]	[-1.69]
2013	0.055***	0.010	-0.002	0.005	-0.013	-0.016
	[3.41]	[0.33]	[-0.35]	[0.60]	[-1.00]	[-0.52]
2014	0.072***	0.003	-0.009	0.003		
	[4.36]	[0.09]	[-1.32]	[0.31]		
Period dummies						
2	-0.014	0.050*	0.011**	0.005	0.062***	0.019
	[-0.71]	[1.67]	[1.99]	[0.61]	[3.62]	[0.78]
3	-0.018	0.075*	0.021***	0.012		
	[-0.89]	[1.95]	[3.03]	[1.19]		
Physician dummies	Yes	Yes	Yes	Yes	Yes	Yes

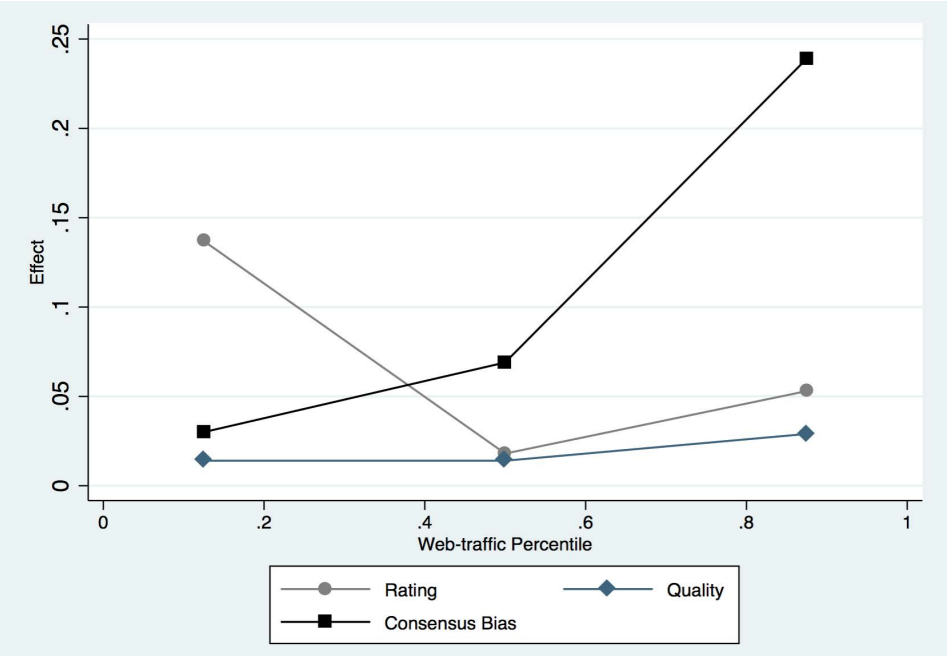
This table presents effect estimates of physician rating disclosure on ratings, quality deductions, and the absolute difference of ratings from prior consensus ratings, with samples partitioned by the length of time (>=1 year or <1 year) that the physicians included in disclosure first met the criterion for inclusion (at least 30 survey responses in a 12 month period) before their ratings were disclosed. °Age dummies for Columns 1-2 and 5-6 are the psychometric set in Newman and Newman [2014], and for Columns 3-4 are the set delineated by CMS (2016) for use in risk adjustment, with one group omitted in each case. Standard errors are clustered at the physician level. *,**,*** denote significance at the .1, .05, and .01 levels respectively. Rating N (>= 1 year) = 97,507 | (< 1 year) = 13,755 , Quality deduction N (>= 1 year) = 33,565 | (< 1 year) = 12,809, Absolute difference N (>= 1 year) = 35,409 | (>= 1 year) = 4,939

TABLE 9: ESTIMATED EFFECTS OF PHYSICIAN RATING DISCLOSURE PARTITIONED BY PRIOR MONTH WEB TRAFFIC TO DISCLOSED INFORMATION

	(1) Rating	(2) Quality deduction	(3) Absolute difference
Cutoff at bottom quartile			
Lower partition	0.137***,†††	-0.014	0.030
	(0.040)	(0.009)	(0.042)
Upper partition	0.016	-0.009	-0.056***,††
	(0.015)	(0.006)	(0.009)
χ^2 test z score	[2.83]	[-0.46]	[2.00]
Cutoff at median			
Lower partition	0.058**	-0.008***	0.015
	(0.023)	(0.004)	(0.031)
Upper partition	0.034	-0.012	-0.057***,††
	(0.028)	(0.011)	(0.015)
χ^2 test z score	[0.66]	[0.34]	[2.09]
Cutoff at top quartile			
Lower partition	0.042**	-0.015***	-0.056
	(0.016)	(0.004)	(0.022)
Upper partition	0.053	-0.029***,††	-0.239***,†††
	(0.043)	(0.004)	(0.021)
χ^2 test z score	[-0.23]	[2.47]	[5.95]

This table presents effect estimates of physician rating disclosure on ratings, quality deductions, and the absolute difference of ratings from prior consensus ratings, with samples partitioned at the physician-month level by one-calendar-month-lagged web traffic to the disclosed information. The estimates for Column 1, 2, and 3 are from the models specified with full controls in Table 3, 4, and 5, respectively. Below each coefficient is the corresponding standard error. Below each pair of effect estimates is a z score reported from the $\chi 2$ test of whether the lower partition effect estimate is significantly more positive than the corresponding upper partition effect estimate. Standard errors are clustered at the physician level. *,**,*** denote the estimate's significance at the .1, .05, and .01 levels, respectively. †,††,††† denote significant results for the $\chi 2$ tests at the .1, .05, and .01 levels, respectively, and are displayed next to the estimate of greatest magnitude included in the corresponding test. Significant results displayed from $\chi 2$ tests remain significant at at least the .1 level, and are either significant at the same level or one level greater or less, after partitioning by 3-calendar-month and 6-calendar-month-lagged web traffic as opposed to 1-calendar-month-lagged web traffic.

FIGURE 6: EFFECTS PARTITIONED BY WEB-TRAFFIC PERCENTILE



This figure displays effect estimates for rating, quality, and consensus bias, partitioned by the physician's percentile rank by prior-calendar-month web traffic to his or her disclosed ratings. "Rating" displays the estimated effect of rating disclosure on *rating* in each partition. "Consensus bias" displays the positive-signed estimated effect of rating disclosure on *absolute difference* in each partition. "Rating" and "consensus bias" are both within the 1-5 rating scale. "Quality" displays the positive-signed estimated effect of rating disclosure on *quality deduction* in each partition, and is within the 0-100 range of the percent of a physician's procedures with no *quality deduction*.

improvement. The tests of H5 and H6 show evidence consistent with that reasoning. Consensus bias, which would make lifting a consumer rating difficult, appears more sensitive to increases in web traffic past low levels than does objective quality improvement. Those low levels are the only area in which the rating-improvement effect declines with increased web traffic.

Table 9 Column 2 shows results of the *quality deduction* analysis partitioned by recent public attention. The results support H5, that the effect of patient satisfaction disclosure on objectively measured quality improvement is greater following greater public attention. The correlation appears only beyond the top-quartile of web traffic, though. This suggests that physicians only realize greater public attention once it reaches high levels, or that it they are aware but not incrementally incentivized to improve as manifested by objective quality measures until the public attention reaches high levels.

Table 9 Column 3 shows results of the *absolute difference* analysis partitioned by recent public attention. All tested cutoffs for partitioning allow rejecting the null in favor of H6, that consensus bias is greater under greater recent public attention.[24] The results show steady increases in consensus bias with greater recent public attention, beginning from the bottom quartile of web traffic. Figure 6 is a graph of the effect estimates for rating, quality, and consensus bias vis-à-vis recent public attention.

2.6. CONCLUSION

This chapter provides evidence of a theoretical tradeoff of publicly disclosing consumer ratings whereby real and positive performance effects persist despite the disclosure creating a bias among raters. Specifically, a health care system's disclosure of patient ratings of its physicians

[24] The models include *rating trend* in order to estimate consensus bias for each partition of web traffic holding *rating trend* constant.

elicited performance improvement by ratings and by objectively measured quality, but generated a bias in the ratings toward the given physician's published consensus. The rating improvement effect was weaker, the objective quality improvement stronger, and the consensus bias effect stronger following greater recent public attention to a physician's disclosed ratings.

Providing this evidence makes three main contributions. First, it forwards research on performance effects of disclosure. The results establish consumer-rating disclosure as a means of driving performance. They also show that performance effects are able to persist in spite of a bias that draws ratings toward prior published consensus ratings. In the case of objectively measured performance, the results are consistent with public attention strengthening disclosure's performance effects. I also help to extend theory regarding that relationship by showing evidence consistent with public attention to consumer-rating disclosure strengthening consensus bias and relatedly impeding rating improvement.

Second, the results are particularly relevant to research on health care disclosure. Long-standing variants of health care disclosure have shown relatively little influence on consumer markets and performance effects. Physician-rating disclosure is relatively recent and is spreading. I find economically and statistically significant positive performance effects of this type of disclosure. This chapter's public attention analysis also provides some of the first direct evidence to test the suggestion that health care disclosure's performance effects are greater under greater public attention to the disclosure. The results support that notion, albeit only in the case of objectively measured performance.

Third, and finally, by establishing evidence of consensus bias, I inform the interpretation and application of consumer ratings. Consumer ratings are increasingly publicly visible. Noting a bias in ratings toward published consensus, and that the bias is stronger when web traffic is

greater, informs the use of ratings to measure underlying service. In particular, consensus bias offers reason for managers to interpret a deviation of subsequent ratings from the published consensus as a dampened signal of recent trends in service. Managers who adjust for the signal dampening would be better able to use consumer ratings to infer service, which is of value in tracing the effects of service on financial performance and in effectively evaluating and rewarding employees.

CHAPTER 3

PERFORMANCE EFFECTS OF SETTING A HIGH REFERENCE POINT FOR PEER-PERFORMANCE COMPARISON

3.1 INTRODUCTION

This study addresses performance effects of setting a reference point for peer comparison above peer-median performance. Providing individuals with relative performance information (RPI), or information for comparing one's own performance to that of one's peers, elicits performance improvement in a variety of compensation settings. These include settings in which performance is not linked to pay or made visible to others (Allcott [2011], Hannan, Krishnan, and Newman [2008], Tafkov [2013]). Theories of social comparison, reference points, expectancy, and goals could guide inquiry into the performance effects of the height of reference points for peer-performance comparison that are commonly displayed along with RPI.[24, 25] Such inquiry could inform the many government, non-profit, and corporate administrators who are using RPI reference points to influence constituents' performance.[26] However, evidence of performance effects of RPI reference point height is lacking.

We provide such evidence through a field experiment in online education. We compare the performance effects of providing the peer top quartile as opposed to the peer median as a reference point for peer comparison within RPI. Each RPI display includes a reference point correctly labeled as one of those two alternatives. The data include measures of a range of

[24] Examples of RPI that include reference points are Gorman [2015] and Daniels and Miller [2011] in hospitals, Allcott [2011] and Schultz et al. [2007] in energy consumption, Harper et al. [2013] regarding a user-generated-content website, and Blanes i Vidal and Nossol [2011] from wholesale and retail.

[25] We study reference points that are percentiles of the peer performance distribution. We use the term "reference point height" in referring to how high a percentile of peer performance is provided as a reference point.

[26] Allcott [2011] and Blanes i Vidal and Nossol [2011] are examples from corporations, Gorman [2015] and Harper et al. [2013] from non-profits, and Kettle et al. [2015] and Hallsworth, List, Metcalfe and Vlaev [2014] from governments.

actions taken in online courses, as well as a log of each instance of a student who receives RPI accessing it, over multiple months. The experimental setting and intervention do not involve explicit incentives for performance, allowing for the identification of the distinct information effect of RPI reference point height. In identifying this effect, we contribute to accounting literature on how RPI affects performance (Murthy [2010], Hannan et al. [2008, 2013], Tafkov [2013]). We also contribute to economic literature on reference points for peer comparison by looking at reference points set through anonymous reports rather than through an introduction to an identifiable peer (Hanushek et al. [2003], Lavy, Silva, and Weinhardt [2012]).

Multiple disciplines offer insight into how RPI and reference points increase performance apart from the rate at which performance is compensated. First, social comparison theory applied in accounting research asserts that RPI facilitates a reward for performance in the form of favorable comparison to one's peers (Brown et al. [2007], Garcia and Tor [2007], Tafkov [2013]). Second, economic research shows that providing higher reference points for total pay increases willingness to exert effort at a given piece rate, lending evidence to reference-dependent preferences for effort provision (Abeler et al. [2011]). Third, RPI reference points that are portrayed as standards for success, as in our study, may exhibit characteristics of goals. Goal theory states that goals energize and focus effort so as to increase performance (Locke and Latham [2002]). Finally, expectancy theory notes that belief in the attainability of a performance level reinforces motivation to work toward it (Atkinson [1957]). We draw from the referenced theories in predicting the effect of providing a higher RPI reference point than the median.[27]

Our predictions and tests add to the insight from a growing number of field studies that address the performance effects of providing RPI reference points (Azmat and Iriberri [2010],

[27] Festinger [1954] and Smith [2000] address foundational theory regarding social comparison, Kahneman and Tversky [1979] regarding reference points and loss-aversion, Locke and Latham [2002] regarding goals, and Atkinson [1957] and Vroom [1964] regarding expectancy-based motivation.

Allcott [2011], Harper et al. [2013], Schultz [2007]). These studies report heterogeneity in performance effects of RPI reference points: individuals underperforming the RPI reference point exhibit a positive performance effect, while those outperforming are less positively or even negatively affected. These studies do not, though, test the performance effects of varying the height of the RPI reference point provided. Theory suggests that providing the higher of two reference points will most positively affect the performance of individuals whose performance initially lies between the two. This implies a concave relationship between the positive performance effect of providing a high reference point and an individual's initial performance. The predicted concavity includes both negative and positive effects in partitions of initial performance, and so we do not ex-ante predict the sign of the average effect.

We find the predicted concave relationship. Relative to displaying the median reference point, displaying the top-quartile reference point negatively affects the performance of initially below-median performers and positively affects the performance of initially 50^{th}-75^{th} percentile performers. We find that the effect among initially top quartile performers depends on the measure of performance. In our experiments, we show students either the outcome-based measure Grade, or the process-based measure Activity Level. Grade is the percent of course problems correctly answered. Activity Level is a weighted sum of course actions such as video views and discussion forum posts. In the case of Grade, surveyed interest in outperforming peers persists at high levels, and top-quartile performers improve more when shown the top-quartile reference point. In the case of Activity Level, interest in outperforming peers is weaker at higher levels, and top-quartile performers improve less when shown the top-quartile reference point.

A few analyses shed further light on the effect of a high reference point for peer-performance comparison. A survey shows that individuals are significantly less confident in their ability to

reach the higher, rather than lower, reference point. Combined with expectancy theory, this data suggests that the negative performance effect when below-median performers view a relatively high reference point is partially due to self-doubt. In terms of demographics, we assess gender as an effect moderator. Prior research shows gender differences in the performance effect of peer comparison when the comparison occurs through an introduction to an identifiable, high performing peer (Eagly [1978], Cross and Madson [1997]). These studies find that women exhibit more positive performance responses to such comparison, and suggest that the result arises from women being more prone to cooperate with and learn from the high-performing peer. We test whether the same result holds when comparison occurs through viewing a percentile of peer performance in a graphical display. In our setting, we do not find that gender moderates the performance effect, consistent with the moderating effect depending on introduction to a peer whom one can choose to cooperate with.

Our study makes three main contributions. First, we show effects of RPI reference point height that operate through the private display of anonymous performance information. This speaks to the growing body of economic, psychology, accounting, and management research on such performance information display as a tool for influencing performance and behavior. This research spans the private and public sector, with outcomes including retail service, educational attainment, energy consumption, web-site content contribution, and taxpaying.[28] Research on the importance of the reference point displayed along with RPI could accordingly have significant policy implications for a variety of corporate and other societal settings.

[28] See Blanes i Vidal and Nossol [2011] for evidence from the retail and wholesale industry, Azmat and Iriberri [2010] regarding education, Allcott [2011] and Schultz et al. [2007] regarding energy consumption, Harper et al. [2013] regarding web-site contributions, and Hallsworth et al. [2014] regarding taxpaying.

Second, analysis of the effects of RPI reference points in isolation informs theory and empirical work in a variety of other accounting and economic streams of research.[29] Research on RPI-related accounting and economic mechanisms might draw from the current paper's results in a number of ways. For example, we find weaker returns to reference point height in the presence of RPI than have been shown for targets of similar height when RPI is hidden (Erez, Early, and Hulin [1985], Locke and Latham [2002]). This offers a partial explanation for the prevalence of easy targets given that targets sometimes derive from or communicate RPI (Aranda et al. [2014], Fisher, Peffer, and Sprinkle [2003], Merchant and Manzoni [1989]).[30] Also, tournament literature raises the problem of motivating those who are very far below or above a rewarded cutoff (Asch [1990], Casas-Arce and Martinez-Jerez [2009]). Our analysis shows an alternative means of performance management for these parts of the performance distribution—providing a lower reference point to low performers, and, in the case of outcome-based performance, a higher reference point to high performers. Further, evidence on the performance returns to reference point height could inform predictions of the effects of supervisor discretion in setting targets that communicate RPI (Bol, et al. [2010]). Lastly, RPI reference points, especially the median and top quartile, are commonly used in measuring corporate performance and evaluating employees.[31] Research in those settings can use our results to account for behavioral responses to the display of these two RPI reference points.

[29] For examples of such RPI use, see Aranda, Arellano, and Davila [2014], Bol et al. [2010], and Murphy [2000] regarding target setting, and Securities and Exchange Commission [2015] regarding disclosure and monitoring, and Gibbons and Roberts [2013] regarding contracting.

[30] See Merchant and Manzoni [1989] for a description of the contrast between theory in favor of targets with less than 50% chance of being attained and the prevalence of much more frequently attainable targets found in practice. Bouwens and Kroos [2011], and Leone and Rock [2002] also find frequently attainable targets in practice.

[31] See Bebchuk and Fried [2005], Bizjak, Lemmon, and Nguyen [2011], and Securities and Exchange Commission [2015] regarding comparison of executive compensation and corporate performance to peer group percentiles, and Berger, Harbring, and Sliwka [2013] and Grote [2005] regarding employee evaluation involving peer group percentiles.

Our third contribution is to the substantial amount of accounting research that shows that the format of information display influences decisions ranging from stock trading to assigning employee bonuses (Bloomfield, Nelson, and Smith [2006], Dilla and Steinbart [2005], Maines and McDaniel [2000]). Our study extends this research by showing how the height of the reference point included in RPI influences performance. Given the common use of RPI and associated percentiles in information displays, the height of reference points set for the purpose of peer-performance comparison is a salient feature to provide evidence on (Song et al. [2015], Bizjak et al. [2011], Gibbons and Roberts [2012] p. 67).

3.2 THEORY AND HYPOTHESIS DEVELOPMENT

Economics-based research addresses multiple functions of RPI that account for its widespread use (Gibbons and Roberts [2012] p. 67). Incorporating RPI in incentive contracts reduces the risk imposed on agents by filtering out common noise from a performance measure linked to incentives (Holmstrom [1982], Lazear and Rosen [1981]). The performance measure is then a more precise signal of effort, and attaching incentives to it imposes less risk on the agent in the form of uncontrollable events that also influence the measure (Banker and Datar [1989]). RPI is also of use to agents in forming expectations of pay for marginal effort in nonlinear incentive schemes. When pay is contingent upon reaching certain levels of relative performance, agents can use information on their proximity to those levels to form such expectations. Empirical studies show that agents subject to nonlinear-incentive schemes expend effort according to their updated rational expectations of pay for marginal effort (Asch [1990], Hannan et al. [2008], Casas-Arce and Martinez-Jerez [2009]).

An emerging stream of accounting and economic research explores an additional role of RPI. Agents incorporate the information into decisions about effort provision even when it does not provide information on the rate of pay for marginal effort (Azmat and Iriberri [2010], Hannan et al. [2008], Murthy [2010], Tafkov [2013]). These studies show generally positive performance effects of providing RPI in fixed-wage, individual-performance piece-rate pay, and in no-pay contexts. In many studied applications, RPI includes reference points for peer comparison, oftentimes peer-median performance (Blanes i Vidal and Nossol [2011], Harper et al. [2013]). Recent economic studies have incorporated reference points as a component of models of utility that traditionally only weigh monetary pay-off and cost of effort. In positively weighting a reference point, utility functions account for "reference-dependent preferences" (Abeler et al. [2011], Farber [2008]). Empirical results provide evidence that reference points for total pay and for peer-performance comparison influence effort provision above and beyond the rate of pay for marginal task performance. Although RPI and reference points both exhibit power in motivating effort, little is known regarding the effect of RPI combined with reference points of varying heights.

A number of field studies report generally positive effects of RPI performance display, but do not test the performance effects of varying the height of included reference points (Allcott [2011], Allcott and Rogers [2014], Azmat and Iriberri [2010], Harper et al. [2013], Schultz et al. [2007]). These studies show heterogeneous performance effects that depend on initial performance relative to the reference point; individuals initially underperforming a reference point exhibit a more positive effect than those initially outperforming it. Proffered explanations include the difficulty of achieving beyond an already high level, as well as a downward psychological attractive power of reference points for those performing above them (Allcott

[2011]), Schultz [2007]). A potential implication of the observed heterogeneity is that a higher reference point, relative to which individuals would be situated differently, would yield different performance effects.

Research on social comparison, reference points, goals, and expectancy all help in understanding how RPI and reference points influence performance. We address each of these streams of research to provide context for our prediction of the performance effects of displaying a relatively high RPI reference point.

Social comparison theory explains how displaying RPI creates performance incentives. RPI allows peer comparison and activates the associated incentives to outperform peers and attain a more positive self-image (Smith [2000], Brown et al. [2007]). Empirical evidence shows that RPI drives performance when peers are identifiable and when they are anonymous, in the presence and absence of performance-based pay, and when one's performance is and is not visible to others (Klar and Giladi [1998], Hannan et al. [2008], Murthy [2010], Tafkov [2013], Xiao and Lucking [2008]).

Behavioral economic research offers insight into the role of reference points—subtly implied numbers that individuals positively weight in economic decisions—in influencing effort. In particular, providing a reference point for expectations of total pay lifts effort toward the level necessary to achieve the given level of pay (Abeler et al. [2011]). Individuals anticipate feelings of loss aversion from receiving pay below the reference point. Utility in the form of a reduced negative deviation from the reference point acts as a performance incentive (Abeler et al. [2011], Farber [2008]). Reference points for effort provision come in forms other than levels of pay. For example, recent research on marathons finds that runners feel a sense of loss from exceeding round number finishing times that serve as reference points (Markle et al. [2015]). Runners exert

effort near the end of the marathon in order to finish a few seconds before the round number time (Allen et al. [2016], Markle et al. [2015]).

The combination of RPI and reference points may also influence performance through a similar mechanism as goals. Goals are explicitly set standards for performance or outcome-achievement, and may be suggested by others or set completely by personal volition. RPI reference points may assume properties of goals to the extent that individuals accept them as standards for success. Goal theory literature has established the ability of assigned goals to improve performance, without rewards for their achievement, through forces including the following: 1) directing attention, 2) energizing activity, 3) affecting persistence, and 4) leading to the arousal, discovery, and/or use of task-relevant capabilities (Locke and Bryan [1969], Locke and Latham [2002]).

Expectancy theory suggests a qualification on the power of RPI reference points to drive performance. It states that motivation to achieve an outcome, such as performing at or above the level of a displayed RPI reference point, depends on perceived attainability of the outcome. Motivation is increasing in "expectancy," or the belief that effort will lead to performance necessary to achieve the outcome (Atkinson [1957], Lawler and Suttle [1973], Vroom [1964]). Goal theory similarly states that belief that a goal is attainable is essential to its motivational effect (Erez and Zidon [1984], Locke et al. [1986], Locke and Latham [2002]). Perceived attainability is particularly relevant to the provision of the RPI reference points addressed in this study, which are inherently unattainable for either a large portion (in the case of median performance) or the majority (in the case of top-quartile performance) of individuals.

An additional motivating force present in our study and in a variety of corporate and public-sector contexts is a visual indication of approval for performing well relative to a peer reference

point (Allcott [2011], Campbell [2002], Vanek Smith [2015]). An example of a visual used in practice is color-coding, with green indicating high and red indicating low performance. Another example is a smiley face for outperforming a reference point. We adopt smiley faces, as used in field experiments from psychology and economics, in order to test the performance effects of RPI reference points when outperforming them is visually congratulated (Allcott [2011], Schultz et al. [2007]).

The development of Hypotheses 2-4, regarding RPI reference point height, take into account the motivating forces of RPI and reference points described. Hypothesis 1 addresses the effect of providing RPI with a reference point in our study to test the intervention's validity as a performance management tool. All of our hypotheses are stated in the alternative form, and are assessed using a two-tailed test.

H1: Providing relative performance information with a congratulated descriptive norm reference point for peer comparison positively affects performance.

We predict a concave relationship between an individual's performance before RPI reference point provision and the performance returns to providing a higher RPI reference point. That hypothesized concave relationship is based in part on our expectation that the top-quartile reference point will not have an incrementally positive effect relative to the median reference point for initially below-median performers. In fact, expectancy theory and forces of reference points and social comparison even suggest the possibility of negative returns to a higher reference point for this group. The higher reference point would impose a lower value on expectancy, a construct positively related to motivation, for individuals performing beneath the

lower reference point (Atkinson [1957]). Also, loss aversion felt through negative deviation from reference points is greater the closer one is to the reference point (Kahneman and Tversky [1979]). Loss aversion is an effort-motivating force that would be weaker with a more distant reference point as long as the individual is similarly interested in surpassing either (Abeler [2011], Heath et al. [1999], Markle et al. [2014]). In terms of utility from congratulations for exceeding an RPI reference point, the lower reference point would offer greater returns to marginal effort. In our study, individuals can anticipate the congratulations using the RPI display legend, which showed a smiley face associated with exceeding the RPI reference point.

From a social comparison theory standpoint, raising the reference point increases the height of the upward social comparison for this group. Social comparison of performance involving great upward distance has been shown to cause discouragement and hurt performance (Rogers and Feller [2015]). Individuals who are below median because they struggle to interact effectively with the course might also resort to ineffective or unsustainable strategies for doing so (Hannan et al. [2008]). Research on tournaments and rank-based-pay also suggests that individuals are discouraged to the point of giving up when they feel that an explicitly rewarded reference point is too high (Bandiera et al. [2013]). A similar force may apply to peer comparison reference points. Further, the upward attractive force of the median reference point may be uniquely powerful due to its relevance to comparison of oneself to what is average given the common desire to consider oneself above-average (Dolan et al. [2012], Larrick et al. [2007]).

On the other hand, goal and reference point research implies some advantage of the higher reference point among initially below-median performers. Goal theory shows that, subject to goal commitment, a higher goal elicits greater performance (Locke and Latham [2002]). Also, individuals who surpass the median might be drawn higher still by the higher reference point.

This concept is in line with utility from eliminating feelings of loss-aversion for performance below a reference point, and with desires to reach a congratulated level of performance (Abeler [2011], Heath et al. [1999], Reno, Cialdini, and Kallgren [1993]). Although theoretical implications for the effect of providing the top-quartile as opposed to median reference point for individuals initially below both are mixed, we predict a slightly negative performance effect among these individuals.

In the partition between the two alternative reference points, forces from goal theory and expectancy theory are more aligned with a positive effect of a higher reference point than in the case of below-median performers. The 50^{th}-75^{th} percentile performers are more likely to view the higher reference point as attainable, and therefore to be committed to it as a goal (Locke and Latham [2002]). In terms of expectancy theory, the belief that effort will lead to performance necessary to reach the reference point would be higher for these individuals than for those who are below median (Atkinson [1957]). Also, for 50^{th}-75^{th} percentile performers, the higher reference point introduces the motivating force of "valance," or satisfaction from reaching a rewarded level of performance (Atkinson [1957], Lawler [1968]). The lower reference point, by contrast, offers valence from rising above a rewarded performance level only to those in this partition of initial performance who fall below the median. Other than potentially greater innate interest in and resulting peer-comparison engagement from the median reference point, theory suggests a uniformly positive effect of the providing the top quartile point instead to the 50^{th}-75^{th} percentile performers (Dolan et al. [2012], Larrick et al. [2007]).

Individuals in the top quartile of performance view a reference point below their position in the distribution in both the case of median and top-quartile RPI reference point display. Neither the top-quartile nor the median reference point, then, offers valence in the form of changing

one's initial performance relative to the RPI reference point. A few qualified forces might work to the positive performance effect of providing the higher reference point to this group. Top-quartile performers are more likely to fall beneath the higher than the lower reference point. The greater prospect and instance of that event when the reference point is higher might yield stronger performance incentives. Also, to the extent that peer-comparison reference points carry downward attractive power, a higher reference point might act as a bulwark to mitigate a resulting performance decline (Schultz et al. [2007]). However, evidence of a downward attractive force of RPI reference points, though, is mixed (Allcott [2011], Harper et al. [2013]). Further, greater innate interest in the median may drive social comparison and related performance among top-quartile performers (Dolan et al. [2012], Larrick et al. [2007]). Finally, the higher reference point reveals to top-quartile performers that they are toward that positive tail of the distribution. This could create concern that one's behavior is economically suboptimal and discourage effort (Schultz et al. [2007]). We predict a positive effect of providing the top-quartile as opposed to the median reference point among initially top-quartile performers despite the noted limitations of forces working toward that effect. We predict that performers initially between the median and top-quartile, those theoretically more uniformly benefitted by viewing the top-quartile reference point, will exhibit greater positive effects from its display than will performers initially in the top quartile.

Hypotheses 2a-c predict the effect of providing the top-quartile as opposed to median reference point in each partition of initial performance addressed above. These predictions include effect directions despite the noted possibility of forces working in the opposing direction. Identifying the direction and significance of effects among partitions helps in understanding the nature of the concavity that Hypotheses 3a-c outline. Identifying the nature of any observed

concavity, in turn, deepens understanding of the overall effect of the higher reference point, which Hypothesis 4 addresses.

H2a: Presenting the peer top quartile, as opposed to the peer median, as a congratulated reference point for performance negatively affects the performance of individuals who are initially below both reference points.

H2b: Presenting the peer top quartile, as opposed to the peer median, as a congratulated reference point for performance positively affects the performance of individuals who are initially between both reference points.

H2c: Presenting the peer top quartile, as opposed to the peer median, as a congratulated reference point for performance positively affects the performance of individuals who are initially above both reference points.

H3a-c address a concave relationship between initial performance and returns to a higher reference point, which H2a-c collectively imply. H3a predicts a relatively more positive effect of the top-quartile as opposed to the median reference point in the partition of performers initially between these alternative reference points than in the outer two partitions. H3b (H3c) predict a more positive effect in the in-between partition than in the lower (higher), helping to further define the shape of performance returns to a higher reference point along the scale of initial performance.

H3a: The performance effect of presenting the peer top-quartile, as opposed to the peer median, as a congratulated reference point for performance is more positive for those who are initially between median and top-quartile performance than for those who are not.

H3b: The performance effect of presenting the peer top-quartile, as opposed to the peer median, as a congratulated reference point for performance is more positive for those who are initially between median and top-quartile performance than for those who are initially below median performance.

H3c: The performance effect of presenting the peer top-quartile, as opposed to the peer median, as a congratulated reference point for performance is more positive for those who are initially between median and top-quartile performance than for those who are initially above top-quartile performance.

H4 addresses the average effect of providing the top-quartile as opposed to median reference point. H2a predicted a negative effect among the half of individuals initially below median. H2b predicted a positive effect among the quarter of individuals who are initially between alternative reference points. H2c and H3c collectively predicted a positive effect for the quarter of individuals above both alternative reference points that is less positive than it is for the quarter of individuals initially in-between the alternatives. H2a-c and H3a-c might involve effects in segments of initial performance that balance out, or alternatively that yield a directional net effect. We test for a net effect, but do not predict its direction given the directional opposition of predicted effects along the scale of initial performance.

H4: Presenting the peer top quartile, as opposed to the peer median, as a congratulated reference point affects performance.

3.3 FIELD SETTING AND DATA

3.3.1 Field Setting

In 2012, MIT and Harvard University jointly founded edX, a nonprofit organization offering free online courses, assessments, and certificates for higher-education courses. HarvardX, our study's field site, is the constituent organization of edX that offers courses from Harvard University faculty members. Enrollment is open globally and with no prerequisites or application process. All instruction occurs online. Course topics range from literature to statistics, and are open for periods ranging from a few weeks to a full year. We conducted experiments in four statistics courses that ranged in enrollment from roughly 6,000 to 25,000.

3.3.2 Experiment Design

Our study's 1x3 experiment design consisted of a control group, which received no RPI display, and two treatment groups, one that received an RPI display with the peer median reference point, and one that received an RPI display with the peer top-quartile reference point. The reference point in a display was labeled either "classmate median" or "classmate top quartile" to correctly reflect the point in the peer distribution that it represented. A student viewing a display was thereby informed which of those two standards he or she was being compared to. We provided students access for a period of over two months to a personal RPI display that included a reference point for peer comparison: either median or top-quartile performance. We delivered the RPI to each of the two treatment groups using weekly emails with a link to the RPI data displays.

To increase exposure to the intervention, we placed links reading, "Check your progress" within the course platform. Control group students clicking the link were directed to the default HarvardX progress chart for the course, showing completion status of individual assignments with no RPI. Treatment group students clicking the link were directed their personal RPI display, below which sat a link to the default HarvardX progress chart for the course. The displays were updated and available daily. Appendix D contains images of the graphs and survey instruments.

In the main experiment we used "Activity Level" (defined in Appendix C and described in section *3.3.3 Data*) as opposed to grade as the measure of performance. The choice of the latter was a matter of data availability; edX could not provide us daily access to grades at the time our proposal was approved, and in some courses approved for the experiment, grades were self-assigned and so not a verifiable indicator of performance. However, this study's referee observed that individuals may respond differently to RPI regarding effort, which Activity Level captures, than they do to RPI regarding an outcome, such as grade. We developed the technology necessary to acquire daily access to objective grades. We ran the experiment both as registered for the conference, with Activity Level as the dependent variable, and as advised by the referee, with grade as the dependent variable. We refer to the former as our "main experiment," and the latter as the "supplemental experiment."

3.3.3 Data

Our study benefits from intricate student-course-level data. Quantitative data include each student's number of clicks on course content, number of days on which they were active, number of video views, number of discussion forum posts, grade, and several other measures of activity in the course. Qualitative data include student demographics and responses to surveys, which we incorporate in the additional analyses.

Activity Level is an aggregate measure of how active the student is in the course. It is a weighted sum of activities that represent course engagement: accessing the course, clicking on course content, watching videos, interacting in the discussion forum, and attempting problems. The weight applied to each summand approximately scales the summand's historical mean to the historical mean of video views. The historical means were measured using data from past iterations of the host courses.

Grade represents the percent of the course's total problems that the student has answered correctly. Problems can be completed asynchronously. This facilitates a flexible, modular learning environment that students can use to suit their particular educational needs. The low mean Grade for students who have attempted problems does not reflect low accuracy, but rather students selecting which material they will complete and having less-than-perfect accuracy in completing the material. Success by Grade in our setting could be thought of as similar to Academy Awards. A director's number of awards is a function both of the number of movies the director chooses to make and the director's success rate in making awarded movies.

Appendix C contains a full list of variable definitions. The dependent variable in the main experiment is Δ Activity Level, and in the supplemental experiment is Δ Grade.[32] The other variables are used in the same manner across both the main and supplemental experiments.

3.4 ANALYSIS

3.4.1 Analytical Approach

[32] We calculate Activity Level and Grade for both courses, but only calculate Δ Activity Level and Δ Grade for the respective experiments wherein each is the dependent variable. This is due to limitations on the longitudinal history in our raw data source for each when it is not the dependent variable. It is outside the scope of this paper to show how Δ Activity Level and Δ Grade correlate during our experiment. Including Activity Level and Grade in descriptive statistics, though, shows how the two correlate over the span of a course.

To make best use of the data, including null results, our study draws both from Null Hypothesis Significance Testing (NHST) and Bayesian analysis. We conduct NHST using ordinary least squares regressions for each hypothesis. When the data fail to reject a null hypothesis, or when our alternative hypothesis predicts no relation, we conduct Bayesian analysis indicating how much more probable we can expect a significant relation to be than we could before the realization of the data.

In selecting our sample, we follow precedent from past economic research in similar field settings, and apply guidance from a methodology study on experiments in online courses. One empirical challenge is that of zero-inflation from the large percentage of individuals who enroll and then do not participate in these courses (Lamb et al. [2010]). Enrolling in a course is free, so individuals often enroll in a noncommittal manner. We exclude from our study students who enroll in, but do not access, the course. Of those who access the course, a large number do not try graded content. We exclude these individuals in the supplemental experiment to avoid zero-inflating the displayed standard of performance. A similar restriction of a sample to the portion of active online community members can be seen in Harper et al. [2013]. This provides individuals with a comparison to others that they deem as similar; such similarity is key to engagement in social comparison and the related effectiveness of RPI (Harper et al. [2013], Tafkov [2013]). Another empirical challenge is the influence of a small number of outliers who utilize an online course approximately ten times more than the 99th percentile (Lamb et al., [2010]). We winsorize values for Activity Level and its component Problem Attempts (number of problems attempted) in their 99th percentiles before using either as a dependent variable to ensure that a small number of extreme outliers do not drive or offset results. This is not necessary for Grade, which is inherently capped at 100. We cluster standard errors at the student level to

correct for auto-correlation from students who enroll in more than one course hosting the experiment. Any students present in more than one course hosting the experiment are included in the same experimental group in all courses, and their experimental group membership was, as with all students, set through random assignment.

3.4.2 Descriptive Statistics

Tables 10 and 11 show the sample selection and descriptive statistics for the main and supplemental experiments, respectively. The courses attract individuals who are on average in their 30's. The majority are male, have at least a bachelor's degree, and live in a developed country. Of those who responded to the pre-course survey, the average student is somewhat to very familiar with the course content and intends to complete at least some course content. The baseline means for Activity Level and Grade are 96.5 and 19.02, allowing students an opportunity to reveal their level of initial performance before the experiment began. Activity Level rises by 17.56 and Grade by 1.98 in the control groups during the main and supplemental experiments, respectively, indicating substantial activity during the experiments separate from the display of RPI.

Tables 12 and 13 show the correlation matrices for each experiment. Both Age and Developed Country are positively correlated with our study's measures of performance. Survey responses indicating prior experience with online courses and commitment to complete the course are also generally positively correlated with performance. Female gender is negatively correlated with prior experience and commitment. Likely through those mediators, it is also negatively correlated with performance.

Tables 14 and 15 show the geographic distribution of students. Students are widely dispersed around the globe. Europe and North America are about equally represented. Asia is

Table 10: Sample Selection and Descriptive Statistics for Main Experiment

Panel A: Sample Selection	
Total Enrollment	24,554
Exclude Students who did not Access the Course	9,375
Final Sample	15,179

Panel B: Descriptive Statistics

	N	Mean	Std. Dev.	25%	75%
Gender	13,260	0.32	0.46	0	1
Age	10,013	32.03	9.75	25	37
Level of Education	10,264	0.84	0.36	1	1
Developed Country	11,599	0.62	0.48	0	1
Familiarity With Subject	532	1.46	0.91	1	2
Commitment to Complete Course	512	2.57	0.61	2	3
Number of Online Courses Previously Enrolled-In	503	5.97	4.55	2	5
Number of Online Courses Previously Completed	526	3.33	3.82	0	4
Grade	10,169	9.31	23.67	0	10
Activity Level	15,177	117.73	247.74	5	70
Δ Activity Level	15,177	21.23	89.43	0	5

This table shows the sample selection and descriptive statistics for the main experiment. Activity Level is the displayed measure of performance in RPI. Demographic data are missing for students who did not fill it in when asked in the registration process and within the course. Grade is missing for students whose records were no longer in the course after we completed development of a code for accessing grade data. Grade is only provided in the main experiment for descriptive purposes. "Level of Education" is an indicator variable for an individual holding a bachelor's or higher degree. "Familiarity With Subject" is on an increasing scale of 0-4. "Commitment to Complete Course" is on an increasing scale of 1-3.

Table 11: Sample Selection and Descriptive Statistics for Supplemental Experiment

Total Enrollment	28,057
Exclude Students who did not try Graded Content	23,597
Final Sample	4,460

Panel B: Descriptive Statistics

	N	Mean	Std. Dev.	25%	75%
Gender	3,902	0.30	0.45	0	1
Age	3,772	30.62	28	24	35
Level of Education	3,857	0.81	0.38	1	1
Developed Country	4,367	0.64	0.47	0	1
Familiarity With Subject	2,242	1.57	0.97	1	2
Commitment to Complete Course	2,404	2.16	0.92	2	3
Number of Online Courses Previously Enrolled-In	2,221	3.44	3.75	1	5
Number of Online Courses Previously Completed	2,221	1.86	2.86	0	2
Activity Level	4,460	222.43	230.17	11	354
Problem Attempts	4,460	94.05	122.40	10	146
Δ Problem Attempts	4,460	9.40	36.74	0	0
Problem-Attempt Accuracy	588	0.28	0.33	0.12	0.38
Grade	4,460	21.65	28.95	1	34
Δ Grade	4,460	2.63	11.22	0	0

This table shows the sample selection and descriptive statistics for the supplemental experiment. Grade is the displayed measure of performance in RPI. Demographic data are missing for students who did not fill it in when asked in the registration process and within the course. Level of Education is an indicator variable for an individual holding a bachelor's or higher degree. "Familiarity With Subject" is on an increasing scale of 0-4. "Commitment to Complete Course" is on an increasing scale of 1-3.

Table 12: Correlation of Descriptive Statistics for Main Experiment

		I	II	III	IV	V	VI	VII	VIII	IX	X
I	Gender										
II	Age	-0.101***									
III	Level of Education	0.022**	0.320***								
IV	Developed Country	0.059***	0.166***	0.106***							
V	Familiarity With Subject	-0.192***	0.093*	0.080	0.011						
VI	Commitment to Complete Course	0.004	-0.031	-0.063	-0.126***	0.073					
VII	Num. Online Courses Prev. Enrolled-In	-0.125*	0.288***	-0.016	0.089*	0.228***	0.032				
VIII	Num. Online Courses Prev. Completed	-0.143***	0.325***	0.027	0.088*	0.343***	0.123**	0.727***			
IX	Grade	-0.052***	0.046***	-0.029***	0.052***	0.042	0.082	-0.006	0.070		
X	Activity Level	-0.035***	0.058***	0.005	0.041***	0.031	0.062	-0.010	0.041	0.775***	
XI	Δ Activity Level	-0.015***	0.034***	-0.018*	0.040***	0.060	0.046	-0.021	0.001	0.376***	0.594***

This table shows a correlation matrix for the variables in the main experiment. *,**,*** denote significance at the .1, .05, and .01 levels, respectively.

Table 13: Correlation of Descriptive Statistics for Supplemental Experiment

		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII
I	Gender													
II	Age	-0.128***												
III	Level of Education	0.016	0.344***											
IV	Developed Country	0.074***	0.144***	0.092***										
V	Familiarity With Subject	0.027	0.084***	0.158***	0.066***									
VI	Commitment to Complete Course	-0.013	-0.016	-0.032	-0.086***	0.014								
VII	Num. Online Courses Prev. Enrolled-In	-0.123***	0.211***	0.037*	0.062***	0.054**	-0.015							
VIII	Num. Online Courses Prev. Completed	-0.135***	0.253***	0.025	0.078***	0.082***	0.032	0.789***						
IX	Activity Level	-0.017	0.058***	-0.010	0.027*	0.016	0.094***	0.062***	0.100***					
X	Problem Attempts	-0.022	0.017	0.006	0.018	0.0435**	0.102***	0.054**	0.098***	0.888***				
XI	Δ Problem Attempts	-0.017	0.014	0.029*	0.022	-0.020	0.042**	0.034	0.026	0.292***	0.230***			
XII	Problem-Attempt Accuracy	-0.055	0.029	-0.012	-0.008	-0.013	0.020	0.066	0.103*	0.039	0.003	0.045		
XIII	Grade	-0.045***	0.034***	0.010*	0.050***	0.047**	0.094***	0.102***	0.148***	0.856***	0.766***	0.277***	0.231***	
XIV	Δ Grade	-0.020	0.046***	-0.028*	0.023	-0.020	0.038*	0.042**	0.031	0.271***	0.209***	0.922***	0.275***	0.315***

This table shows a correlation matrix for the variables in the supplemental experiment. *,**,*** denote significance at the .1, .05, and .01 levels, respectively.

Table 14: Number of Students by Continent in Main Experiment

	Number of Students	
Continent	Developed Country	Developing Country
Africa	0	626
Asia	113	2,629
Europe	3,329	1
North America	3,533	554
Oceania	0	78
South America	0	578
	6,975	4,466

This table shows the distribution of individuals in the main experiment across the developed and developing world, subcategorized by continent.

Table 15: Number of Students by Continent in Main Experiment

	Number of Students	
Continent	Developed Country	Developing Country
Africa	0	107
Asia	36	963
Europe	1,277	2
North America	1,398	163
Oceania	114	1
South America	0	306
	2,825	1,542

This table shows the distribution of individuals in the supplemental experiment across the developed and developing world, subcategorized by continent.

close behind at about 80% of the population from Europe. Africa and South America each account for about 10-30% of the better-represented continents.

Tables 16 and 17 show RPI access across treatments. The grade RPI attracts more attention than Activity Level RPI. In both cases, only a small portion of the sample chooses to view the RPI. Future research can explore whether RPI could have a much stronger effect in practice were it delivered with explicit incentives for opening it and thereby viewed more broadly. Our study speaks to settings in which viewing RPI is voluntary and a relatively small number of individuals opt to.

Tables 18 and 19 show the distribution of the dependent variable for each experiment (Δ Activity Level for the main experiment, and Δ Grade for the supplemental experiment) by treatment and level of initial performance. Figures 7 and 8 provide a graphical representation of the distribution of the dependent variable for each experiment.

3.4.3 Results

In both the main and supplemental experiments, we find evidence to support H1 that providing RPI positively affects performance. The effect estimates for Activity Level and Grade appear in Column 1 of Table 20 Panel A and Table 21 Panel A. Both experiments also provide evidence to support H2a, that the lower reference point more positively affects performance than the higher reference point among initially low performers. These results are shown in Column 2 of Table 20 Panel A and Table 21 Panel A. In the main experiment, we find statistically significant evidence in support of H2b. Specifically, the higher reference point more positively affects performance than the lower reference point among individuals initially in between these two reference points. Column 3 of Table 20 Panel A shows this result. We find similar evidence, although not at a statistically significant level, in the supplemental field experiment, shown in Column 3 of Table

Table 16: RPI Access in Main Experiment

Sample	N
RPI	
Never Accessed RPI	9,752
Accessed RPI Once	213
Accessed RPI More than Once	367
RPI_M	
Never Accessed RPI	4,890
Accessed RPI Once	108
Accessed RPI More than Once	176
RPI_T	
Never Accessed RPI	4,862
Accessed RPI Once	105
Accessed RPI More than Once	191

This table describes the distribution of RPI graph access in the main experiment. The table categorizes individuals by experimental condition. The differences in RPI access between the RPI_M and RPI_T groups are not statistically significant.

Table 17: RPI Access in Supplemental Experiment

Sample	N
RPI	
Never Accessed RPI	2,995
Accessed RPI Once	409
Accessed RPI More than Once	372
RPI_M	
Never Accessed RPI	1,503
Accessed RPI Once	203
Accessed RPI More than Once	180
RPI_T	
Never Accessed RPI	1,503
Accessed RPI Once	206
Accessed RPI More than Once	192

This table describes the distribution of RPI graph access in the supplemental experiment. The table categorizes individuals by experimental condition. The differences in RPI access between the RPI_M and RPI_T groups are not statistically significant.

Table 18: Distribution of Δ Activity Level in Main Experiment

Partition	N	Mean Δ Activity Level
Control		
Initially Below Average	2,659	10.78
Initially Third Quartile	1,124	16.02
Initially Top Quartile	1,056	36.27
All	4,839	17.56
RPI_M		
Initially Below Average	2,966	16.43
Initially Third Quartile	1,068	16.48
Initially Top Quartile	1,140	45.33
All	5,174	22.81
RPI_T		
Initially Below Average	2,916	12.54
Initially Third Quartile	1,094	24.96
Initially Top Quartile	1,148	35.00
All	5,158	20.18

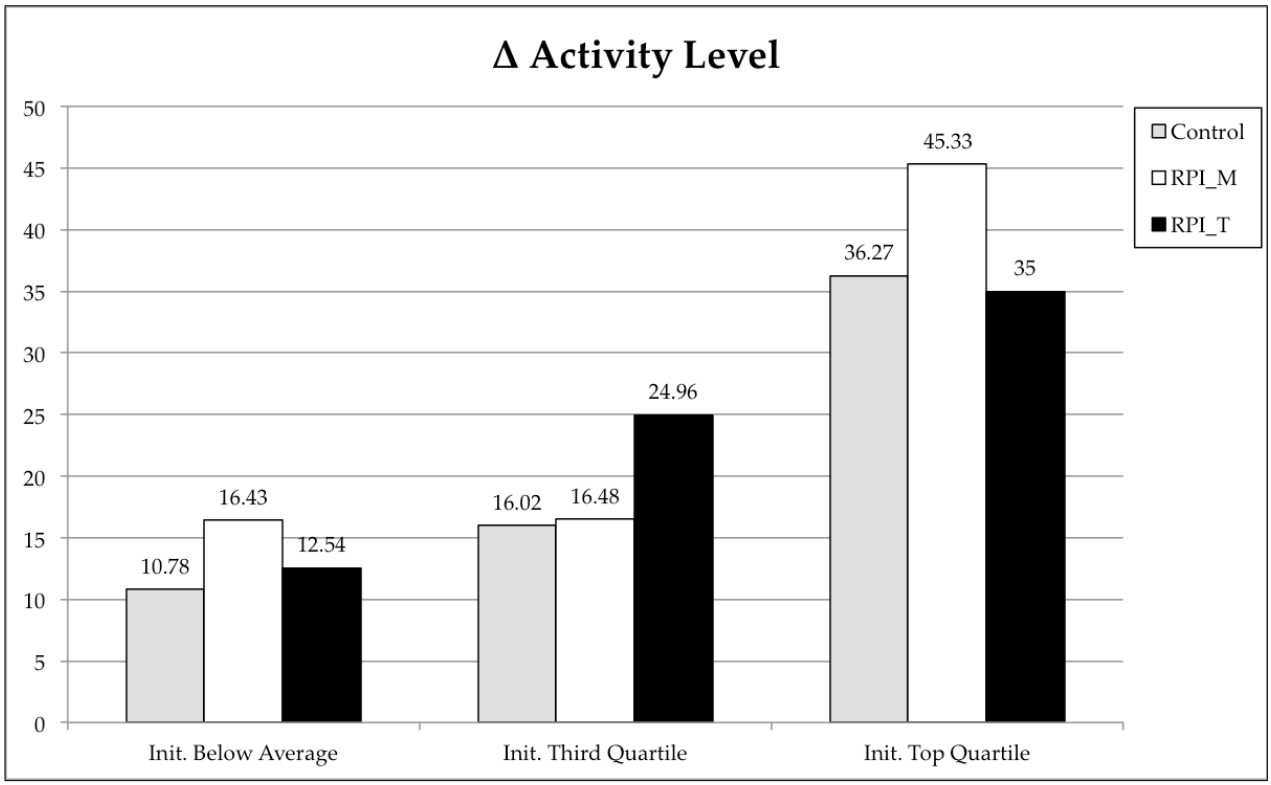
This table shows the mean Δ Activity Level for individuals in the main experiment. The table categorizes individuals by experimental condition and by initial performance level.

Table 19: Distribution of Δ Grade in Supplemental Experiment

Partition	N	Mean Δ Grade
Control		
Initially Below Average	995	2.16
Initially Third Quartile	234	1.94
Initially Top Quartile	236	1.22
All	1,465	1.98
RPI_M		
Initially Below Average	1,032	3.95
Initially Third Quartile	234	1.73
Initially Top Quartile	237	0.65
All	1,503	3.09
RPI_T		
Initially Below Average	1,023	2.95
Initially Third Quartile	233	3.04
Initially Top Quartile	236	1.98
All	1,492	2.81

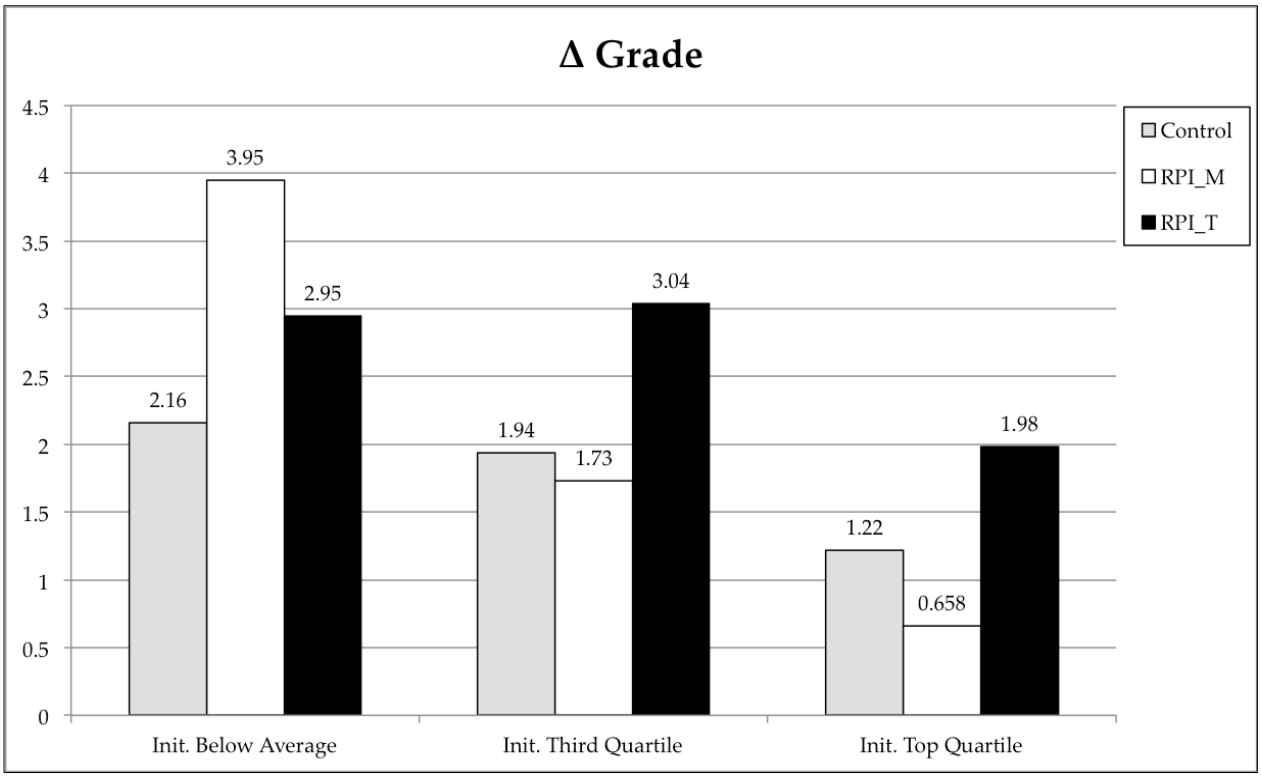
This table shows the mean Δ Grade for individuals in the supplemental experiment. The table categorizes individuals by experimental condition and by initial performance level.

Figure 7: Main Experiment Outcomes by Treatment and Initial Performance



This figure shows the mean Δ Activity Level for each treatment and in each partition of initial performance. These data are from the main experiment, with Activity Level as the displayed measure of performance in RPI.

Figure 8: Supplemental Experiment Outcomes by Treatment and Initial Performance



This figure shows the mean Δ Grade for each treatment and in each partition of initial performance. These data are from the supplemental experiment, with grade as the displayed measure of performance in RPI.

Table 20: RPI and Reference Point Effects in Main Experiment

Panel A. RPI Effect and Reference Point Partitioned Effects				
	1	2	3	4
		Δ Activity Level		
RPI	2.77**			
	[2.46]			
RPI_T		-3.71**	7.43**	-6.57
		[-2.21]	[2.24]	[-1.62]
Course Fixed Effects	Yes	Yes	Yes	Yes
Bayes Factor				20.81
N	15,126	5,882	2,168	2,288
Clustering	Student	Student	Student	Student
Sample	All	RPI & Init. BelowAverage	RPI & Init. Third Quartile	RPI & Init. Top Quartile

This panel shows effect estimates for the main experiment. Column 1 shows the effect of RPI. Columns 2-4 show, for individuals in the RPI condition, the effect of displaying RPI with the top-quartile as opposed to median reference point. These columns are partitioned by an individual's initial level of performance. T-statistics are in brackets. *, **, *** denote statistical significance at the .1, .05, and .01 levels, respectively.

Table 20: RPI and Reference Point Effects in Main Experiment (Continued)

Panel B. Reference Point Effect Concavity and Net Effect				
	1	2	3	4
			Δ Activity Level	
RPI_T	-4.77***	-3.52**	-6.57	-2.14
	[-2.70]	[-2.09]	[-1.62]	[-1.32]
Init. Third Quartile	-3.31	4.49**	-22.61***	
	[-1.63]	[2.19]	[-6.73]	
RPI_T x Init. Third Quartile	12.46***	11.04***	13.69***	
	[3.29]	[2.87]	[2.70]	
Course Fixed Effects	Yes	Yes	Yes	Yes
Bayes Factor				3.79
N	10,332	8,044	4,450	10,332
Clustering	Student	Student	Student	Student
Sample	RPI	RPI & (Init. Below Average or Init. Third Quartile)	RPI & (Init. Third Quartile or Init. Top Quartile)	RPI

This panel shows effect estimates for the main experiment. Columns 1-3 illustrate, with the coefficient on RPI_T x Init. Third Quartile, the concavity in initial performance of providing the top-quartile as oppposed to median reference point. The effect is concave in initial performance in that it is most positive for individuals in the third quartile of initial performance. Column 4 shows the average effect of providing the top-quartile as opposed to median reference point. T-statistics are in brackets. *,**,*** denote statistical significance at the .1, .05, and .01 levels, respectively.

Table 21: RPI and Reference Point Effects from Supplemental Experiment

Panel A. RPI Net Effect and Reference Point Partitioned Effects				
	1	2	3	4
			Δ Grade	
RPI	0.97***			
	[3.06]			
RPI_T		-1.00*	1.31	1.32**
		[-1.67]	[1.41]	[2.14]
Course Fixed Effects	Yes	Yes	Yes	Yes
Bayes Factor			2.04	
N	4,460	2,055	467	473
Sample	All	RPI & Init. Below Average	RPI & Init. Third Quartile	RPI & Init. Top Quartile

This panel shows effect estimates for the supplemental experiment. Column 1 shows the effect of RPI. Columns 2-4 show, for individuals in the RPI condition, the effect of displaying RPI with the the top-quartile as opposed to median reference point. These columns are partitioned by an individual's initial level of performance. T-statistics are in brackets. *, **, *** denote statistical significance at the .1, .05, and .01 levels, respectively.

Table 21: RPI and Reference Point Effects from Supplemental Experiment (Continued)

Panel B. Reference Point Effect Concavity and Net Effect				
	1	2	3	4
			Δ Grade	
RPI_T	-0.56	-1.00*	1.32**	-0.27
	[-1.13]	[-1.67]	[2.14]	[-0.61]
Init. Third Quartile	-1.61**	-2.22***	1.13*	
	[-2.53]	[-3.23]	[1.86]	
RPI_T x Init. Third Quartile	1.88*	2.31**	-0.062	
	[1.77]	[2.09]	[-0.05]	
Bayes Factor			1.93	2.79
N	2,995	2,522	923	2,995
Sample	RPI	RPI & (Init. Below Average or Init. Third Quartile)	RPI & (Init. Third Quartile or Init. Top Quartile)	RPI

This panel shows effect estimates for the supplemental experiment. Columns 1-3 illustrate, with the coefficient on RPI_T x Init. Third Quartile, the concavity in initial performance of the effect of providing the top-quartile rather than median reference point. The effect is concave in initial performance in that it is most positive for individuals in the third quartile of initial performance. Column 4 shows the average effect of providing the top-quartile as opposed to median reference point. T-statistics are in brackets. *,**,*** denote statistical significance at the .1, .05, and .01 levels, respectively.

21 Panel A. The Bayes Factor of 2.04 warrants updating a prior of no relationship in favor of the model's estimate of a positive relationship to a level twice the probability before the realization of the data. Among individuals initially in the top-quartile of performance, the estimate in Column 4 of Table 21 Panel A supports H2c that displaying the higher rather than lower reference point for Grade yields a more positive performance effect. This result does not hold when the performance measure is Activity Level, as is visible in Column 4 of Table 20 Panel A. In fact, the substantial Bayes Factor of over 20 suggests that providing the higher rather than lower RPI reference point yields a negative effect among initially top-quartile performers. Our hypothesis noted reasons why this might be the case, including a lack of desire to be above the top-quartile of peers by a process measure, and even concerns that one is putting in more than the optimal amount of effort. We address this result conceptually in Section 5: Discussion.

The results in each partition of initial performance form the basis for a concave relationship between initial performance and the positive effect of displaying the higher rather than lower reference point as predicted in H3a. In Table 20 Panel B and in Table 21 Panel B, Column 1 shows a concave relationship, captured by the positive coefficient on the interaction of RPI_T and Initially Third Quartile. Third quartile refers to the 50^{th}-75^{th} percentile. Column 2 in each of those two panels reveals that the concavity is in part due to the higher reference point yielding greater performance effects among the initially third quartile performers than among low performers, as predicted in H3b. Column 3 in each of those two panels tests H3c, that the concavity is in part due to the higher reference point yielding greater performance effects among the third quartile than the top quartile of performers. This is the case in the main experiment, with Activity Level as the measure of performance. In the supplemental experiment, with Grade as the measure of performance, the higher reference point does not yield significantly better

performance results among third quartile than top quartile performers. The Bayes Factor of 1.93 for that test indicates that such a relation is roughly twice as likely in light of the data.

Figures 9 and 10 provide a graphical representation of the concave relationship between the performance effects of providing a relatively high reference point and an individual's initial performance. The tendency of high performers to respond with little-to-no competitiveness upon viewing the high reference point for Activity Level produces a hill shape in the main experiment, shown in Figure 9. The tendency of high-performers to respond competitively to viewing the high reference point for Grade produces a plateau shape in the supplemental experiment, shown in Figure 10.

We test for an average performance effect with H4. We do not find an average effect of providing the top-quartile as opposed to median reference point. The tests of this hypothesis are shown in Column 4 of Table 20 Panel B and Table 21 Panel B for the main and supplemental experiment, respectively. The Bayes Factor of 3.79 for the main experiment, with Activity Level as the performance measure, suggests that we can update a prior of no relationship in favor of the model's estimate of a negative effect to a level of 3.79 times the probability before the realization of the data. In the supplemental experiment, with Grade as the performance measure, the Bayes Factor is 2.79 in support of updating a prior of no relationship to that of a negative effect.

3.4.4 Additional Analyses

To illustrate drivers and determinants of the main effects, we present the results of a few additional analyses. The first set of additional analyses addresses differences in performance effects by gender. Studies of social interaction point to differences in the cooperativeness and benefit from interactions dependent on gender (Eagly [1978], Cross and Madson [1997]). Recent

evidence suggests females exhibit more positive performance effects from interaction with high performing peers than do males (Lavy et al. [2008]). We test whether females similarly benefit more from comparison to a high reference point of peer-performance, or whether such gender-based heterogeneity is weaker when the element of social interaction in peer comparison is weaker. Three other background variables that the majority of students provide information on are age, residence in a developed country, and level of education. We include these, as well as original performance, in a fully interacted model with RPI reference point height in testing the moderating effects of gender. The results are shown in Table 22. We do not find that women benefit more than men from the provision of RPI, or from the provision of the relatively higher reference point within RPI. The lack of a differential effect is in line with theory that the gender differential in performance response to peer comparison derives from the associated opportunities to interact and cooperate with an individual revealed to be a high performer, which anonymous reporting does not provide.

In Table 23, we look at the source of the improvement in Grade in the supplemental experiment. The result may arise through a quantity, quality, or joint quantity-quality mechanism. Specifically, RPI may motivate students to attempt more problems, to answer problems with greater accuracy, or some combination of the two. We find that Grade RPI display led to a higher quantity of problems attempted, but not a statistically significant increase in the accuracy of problem attempts. This suggests that the RPI led to increased effort as opposed to aptitude. This result is in line with theories of reference-dependent preferences for effort provision (Abeler et al. [2011]).

A final set of additional analyses draws on a post-experiment survey that measured interest in the alternative reference points and confidence in their attainability. Tables 28 and 29

tabulate the survey responses and related tests. In the case of Activity Level, interest in viewing and outperforming the lower reference point is higher than it is for the higher reference point. Some individuals are interested in outperforming the median, but less interested in outperforming the top quartile. In the case of Grade we found equal interest in viewing the two reference points. We also found interest in outperforming peers that persists in its strength even at levels above the higher reference point.

These results help to understand the drop-off in the performance effect of the higher reference for those who've exceeded it when performance is process based (Activity Level), but not when it is outcome based (Grade). In the case of Activity-Level RPI, the high reference point is less may feel complacent and even wonder whether their level of process-performance is suboptimal. In the latter, the survey evidence suggests desires to outperform the higher reference point, a goal that the top-quartile RPI treatment facilitates by showing them the higher reference point.

In the case of both outcome and process performance, confidence in the achievability of median performance is significantly greater than confidence in the achievability of top-quartile performance. We find that the lower performers, those for whom the top quartile is particularly distant, perform better if shown the relatively lower reference point. This result is in line with the predictions of expectancy theory that motivation to achieve an outcome is increasing in its perceived attainability.

3.4.5 Alternative Model Specification

In analyzing data and receiving workshop feedback, we have come across a more parsimonious display of our hypothesis tests than breaking our sample into several factions for subsample comparisons as we did in Tables 18-21. Given that our tests are comparing means, a cell means

model allows displaying each mean of interest in each subsample in a grid. This is shown in Tables 24 and 26 for the main and supplemental experiments, respectively. F tests then allow comparing means within the model. Finally, X^2 tests allow comparing effect estimates, necessary to test whether the effect of the relatively high reference point is indeed statistically significantly stronger among individuals initially between as opposed to outside the two alternative reference points. The F and X^2 test results are in Tables 25 and 26, which address the main and supplemental experiments, respectively. This specification yields the same rejections of null hypotheses at the same levels of statistical significance as do the specifications in Tables 20-21.

The last additional analysis looks at the effect of RPI regarding a reported measure (Grade) and the correlation between grade and an unreported measure (Activity Level). Table 30 shows a decrease in correlation after reporting grade, as captured by the interaction term on Grade and RPI.

3.5 DISCUSSION

This study most directly contributes to a growing body of economic, psychology, accounting and management literature on the distinct effects of RPI, or those separate from pay for or visibility of relative performance (Blanes i Vidal and Nossol [2011], Hannan et al. [2008], Harper et al. [2013], Tafkov [2013]). While such applications of RPI often include a reference point for peer comparison, little, if any, empirical evidence has established the effects of RPI reference point height (Allcott [2011], Harper et al. [2013]).

Principally, we offer some of the first evidence that the performance effect of providing a relatively high RPI reference point depends on initial performance. Our study indicates the

Table 22: RPI and Reference Point Effects by Demographics

	1	2	3	4
	Δ Activity Level		Δ Grade	
RPI	1.00		0.04	
	[0.26]		[0.04]	
RPI_T		8.71*		-2.40
		[1.74]		[-1.46]
Gender	-4.67***	-0.95	-0.20	-0.77
	[2.97]	[-0.47]	[-0.37]	[-1.11]
Level of Education	3.27	0.02	0.94	-0.16
	[1.57]	[0.01]	[1.62]	[-0.16]
Age	0.02	0.30**	-0.02	0.03
	[0.25]	[2.56]	[-0.87]	[0.66]
Developed Country	5.62	5.05***	0.39	-0.72
	[3.67]	[2.63]	[0.71]	[-0.93]
RPI x Gender	2.30		-0.363	
	[1.12]		[-0.48]	
RPI x Level of Education	-2.75		-0.11	
	[-1.03]		[-0.13]	
RPI x Age	0.13		0.03	
	[1.11]		[0.90]	
RPI x Developed Country	-2.12		-0.05	
	[-1.06]		[-0.07]	
RPI_T x Gender		-2.73		0.29
		[-1.02]		[0.29]
RPI_T x Level of Education		0.97		2.17*
		[0.29]		[1.90]
RPI_T x Age		-0.27*		-0.03
		[-1.85]		[-0.71]
RPI_T x Developed Country		-3.08		2.18**
		[-1.20]		[2.20]
N	9,598	6,450	3,611	2,437
Sample	All Main Experiment	RPI Main Experiment	All Supplemental Experiment	RPI Supplemental Experiment

This table shows effect estimates for providing RPI and providing the top-quartile rather than median reference point by the study's demographic variables. The interaction terms in Columns 1-2 illustrate any dependence of the effect of RPI and the higher reference point on Δ Activity Level in the main experiment. The interaction terms in Columns 3-4 illustrate any dependence of the effect of RPI and the higher reference point on Δ Grade in the supplemental experiment. T-statistics are in brackets. *,**,*** denote statistical significance at the .1, .05, and .01 levels, respectively.

Table 23: Effect of Grade RPI on Problem-Attempt Quantity and Accuracy

	1	2
	Δ Problem Attempts	Problem-Attempt Accuracy
RPI	2.28**	3.27
	[2.06]	[1.41]
N	4,460	588
Sample	All	Attempted Problems

This table shows effect estimates for the supplemental experiment of the effect of RPI on problem-attempt quantity and accuracy. Problem-Attempt Quantity is captured by Δ Problem Attempts. Problem-Attempt Accuracy is only calculable for the subsample of students who attempted a problem during the experiment. T-statistics are in brackets. *,**,*** denote statistical significance at the .1, .05, and .01 levels, respectively.

Table 24: Cell Means Model for Main Experiment

	Control	RPI_M	RPI_T
Below Median	c_1 10.78 (0.943)	c_4 16.43 (0.463)	c_7 12.54 (0.969)
Third Quartile	c_2 16.02 (1.815)	c_5 16.48 (1.772)	c_8 24.96 (3.452)
Above Top Quartile	c_3 36.27 (2.514)	c_6 45.33 (3.376)	c_9 35 (2.980)
	$c_{1...3}$ 17.56 (0.934)	$c_{4...6}$ 22.81 (1.289)	$c_{7...9}$ 20.18 (1.207)
		$c_{4...9}$ 21.49 (0.884)	

This table shows cell means for Δ Activity Level from the main experiment. Cells are the nine categories from the matrix of initial performance (Below Median, Third Quartile, Above Top Quartile) and experimental condition (Control, RPI_M, RPI_T). Each cell contains a coefficient from an OLS regression on Δ Activity Level of a categorical variable representing an individual's belonging to the cell. Standard errors are in parentheses. All coefficients are statistically significant at the .01 level.

Table 25: Hypothesis Tests for Main Experiment

Cells	Coefficient	Intercept	Hypothesis	Test Statistic	P-Value
$c_{1\ldots3}$	17.56	None	H_0: $c_{1\ldots3} = c_{4\ldots9}$	$F = 5.12^{**}$	0.023
$c_{4\ldots9}$	21.49		H_A: $c_{1\ldots3} < c_{4\ldots9}$		
c_4	16.43	None	H_0: $c_4 = c_7$	$F = 4.79^{**}$	0.028
c_7	12.54		H_A: $c_4 > c_7$		
c_5	16.48	None	H_0: $c_5 = c_8$	$F = 4.84^{**}$	0.027
c_8	24.96		H_A: $c_5 < c_8$		
c_6	45.33	None	H_0: $c_6 = c_9$	$F = 2.45$	0.117
c_9	35.00		H_A: $c_6 < c_9$		
$c_{7\ \&\ 9}$	-4.86	$c_{4\ \&\ 6}$	H_0: $c_{7\ \&\ 9} = c_8$	$\chi^2 = 11.13^{***}$	0.000
c_8	7.43	c_5	H_A: $c_{7\ \&\ 9} < c_8$		
c_7	-3.71	c_4	H_0: $c_7 = c_8$	$\chi^2 = 9.09^{***}$	0.002
c_8	7.43	c_5	H_A: $c_7 < c_8$		
c_8	7.43	c_5	H_0: $c_8 = c_9$	$\chi^2 = 7.46^{***}$	0.006
c_9	-6.57	c_6	H_A: $c_8 > c_9$		
$c_{4\ldots6}$	22.81	None	H_0: $c_{4\ldots6} = c_{7\ldots9}$	$F = 1.75$	0.185
$c_{7\ldots9}$	20.18				

This table shows the hypothesis tests for the main experiment, with Δ Activity Level as the dependent variable, based on the cell means model in Table 16. F tests compare coefficients within the cell means model as displayed in Table 16 with the intercept suppressed. $\chi 2$ tests compare coefficients between variants of the cell means model with different cells ommitted to serve as the intercept. The F tests compare the RPI to to the control condition, and the top-quartile to the median reference point condition. The $\chi 2$ tests show concavity in initial performance of the performance effect of providing the top-quartile rather than median reference point.

Table 26: Cell Means Model for Supplemental Experiment

	Control	RPI_M	RPI_T
Below Median	c_1 2.16 (0.297)	c_4 3.95 (0.463)	c_7 2.95 (0.381)
Third Quartile	c_2 1.94 (0.484)	c_5 1.73 (0.509)	c_8 3.04 (0.783)
Above Top Quartile	c_3 1.22 (0.463)	c_6 0.65 (0.304)	c_9 1.98 (0.539)
	$c_{1...3}$ 1.98 (0.225)	$c_{4...6}$ 3.09 (0.332)	$c_{7...9}$ 2.81 (0.300)
		$c_{4...9}$ 2.95 (0.224)	

This table shows cell means for Δ Grade from the supplemental experiment. Cells are the nine categories from the matrix of initial performance (Below Median, Third Quartile, Above Top Quartile) and experimental condition (Control, RPI_M, RPI_T). Each cell contains a coefficient from an OLS regression on Δ Grade of a categorical variable representing an individual's belonging to the cell. Standard errors are in parentheses. All coefficients are statistically significant at the .01 level.

Table 27: Hypothesis Tests for Supplemental Experiment

Cells	Coefficient	Intercept	Hypothesis	Test Statistic	P-Value
$c_{1...3}$	1.98	None	H_0: $c_{1...3} = c_{4...9}$	$F = 9.36$***	0.002
$c_{4...9}$	2.95		H_A: $c_{1...3} < c_{4...9}$		
c_4	3.95	None	H_0: $c_4 = c_7$	$F = 2.79$*	0.095
c_7	2.95		H_A: $c_4 > c_7$		
c_5	1.73	None	H_0: $c_5 = c_8$	$F = 1.98$	0.159
c_8	3.04		H_A: $c_5 < c_8$		
c_6	0.65	None	H_0: $c_6 = c_9$	$F = 4.60$**	0.032
c_9	1.98		H_A: $c_6 < c_9$		
$c_{7\ \&\ 9}$	-0.56	$c_{4\ \&\ 6}$	H_0: $c_{7\ \&\ 9} = c_8$	$\chi^2 = 3.15$*	0.075
c_8	1.31	c_5	H_A: $c_{7\ \&\ 9} < c_8$		
c_7	-1.00	c_4	H_0: $c_7 = c_8$	$\chi^2 = 4.36$**	0.036
c_8	1.31	c_5	H_A: $c_7 < c_8$		
c_8	1.31	c_5	H_0: $c_8 = c_9$	$\chi^2 = 0.00$	0.991
c_9	1.32	c_6	H_A: $c_8 > c_9$		
$c_{4...6}$	3.09	None	H_0: $c_{4...6} = c_{7...9}$	$F = 0.37$	0.542
$c_{7...9}$	2.81				

This table shows the hypothesis tests for the supplemental experiment, with Δ Grade as the dependent variable, based on the cell means model in Table 17. F tests compare coefficients within the cell means model as displayed in Table 17 with the intercept suppressed. $\chi 2$ tests compare coefficients between variants of the cell means model with different cells ommitted to serve as the intercept. The F tests compare the RPI to to the control condition, and the top-quartile to the median reference point condition. The $\chi 2$ tests show concavity in initial performance of the performance effect of providing the top-quartile rather than median reference point.

Table 28: Survey Responses and Wilcoxon Signed-Rank Comparisons of Survey Responses for Main Experiment

Panel A: Survey Questions (the number of students selecting a response sits beside the response in parentheses)

1. Are you interested in seeing how your activity in the course compares to the...			
classmate median:	no (16)	somewhat (20)	yes (21)
classmate top quartile:	no (22)	somewhat (14)	yes (21)
2. How important is it to you to be more active in the course than…			
50% of your classmates:	not at all important (28)	somewhat important (15)	important (16)
75% of your classmates:	not at all important (31)	somewhat important (15)	important (13)
2. How confident are you in your ability to be more active in the course than…			
50% of your classmates:	not at all confident (7)	somewhat confident (20)	confident (29)
75% of your classmates:	not at all confident (11)	somewhat confident (20)	confident (25)

Panel B: Wilcoxon Signed-Rank Comparisons of Survey Responses

Interest in viewing reference point

z score: 2.12 in favor of median reference point
p-val: 0.033
N=57

Importance of reaching reference point

z score: 1.89 in favor of median reference point
p-val: 0.057
N=59

Confidence in ability to reach reference point

z score: 2.82 in favor of median reference point
p-val: 0.004
N=56

This table shows survey questions and responses regarding individuals' opinions of the peer median and top-quartile reference points, as well as a comparison of responses. In comparing responses for each question, the least affirmative response is coded as 1, the intermediate response as 2, and the most affirmative response as 3.

Table 29: Survey Responses and Wilcoxon Signed-Rank Comparisons of Survey Responses for Supplemental Experiment

Panel A: Survey Questions (the number of students selecting a response sits beside the response in parentheses)

1. Are you interested in seeing how your grade in the course compares to the...			
classmate median:	no (6)	somewhat (13)	yes (22)
classmate top quartile:	no (6)	somewhat (13)	yes (22)
2. How important is it to you to get a higher grade in the course than…			
50% of your classmates:	not at all important (13)	somewhat important (10)	important (16)
75% of your classmates:	not at all important (14)	somewhat important (11)	important (14)
2. How confident are you in your ability to get a higher grade in the course than…			
50% of your classmates:	not at all confident (5)	somewhat confident (12)	confident (21)
75% of your classmates:	not at all confident (8)	somewhat confident (15)	confident (16)

Panel B: Wilcoxon Signed-Rank Comparisons of Survey Responses

Interest in viewing reference point

z score: 0.00
p-val: 1.000
N=41

Importance of reaching reference point

z score: 0.70 in favor of median reference point
p-val: 0.479
N=42

Confidence in ability to reach reference point

z score: 2.49 in favor of median reference point
p-val: 0.012
N=39

This table shows survey questions and responses regarding individuals' opinions of the peer median and top-quartile reference points, as well as a comparison of responses. In comparing responses for each question, the least affirmative response is coded as 1, the intermediate response as 2, and the most affirmative response as 3.

Table 30: Effect of Grade RPI on Correlation Between Grade and Activity Level

	1
	Grade at Course End
RPI	10.09***
	[3.51]
Activity Level at Course End	0.14***
	[11.39]
Activity Level at Course End * RPI	-0.104***
	[-3.20]
N	4,460
Sample	All

This table shows effect estimates for the supplemental experiment of the effect of RPI on the correlation between grade and RPI. This effect is captured by the interaction of Grade and RPI. *,**,*** denote statistical significance at the .1, .05, and .01 levels, respectively.

quantiles of performance in which the performance effect of viewing the top-quartile as opposed to median reference point are negative and positive. The effect is negative for initially below-median performers. The effect is positive for individuals initially in between the two reference points. The effect is positive for top-quartile performers in the case of the outcome-based measure Grade. However, it is nearly statistically significantly negative in the case of the process-based measure Activity Level. In developing hypotheses, we noted a few reasons why this might occur, and these seem to apply in the case of Activity Level. One was that individuals might be more interested in checking their comparison to the median than to the top-quartile. Tables 20 and 21 show that individuals report more interest in viewing and comparing favorably to median performance than top-quartile performance. This result holds in untabulated tests in which we restrict the survey to the responses of top-quartile performers. A second, related reason was that individuals might interpret their standing above the higher reference point as a sign of suboptimal behavior. If so, this plausibly applies more to Activity Level than grade given that individuals feel a greater sense of optimality in performance that reveals high skill or intellect (Tafkov, [2013]). Overall, our study shows that performance effects of providing a higher rather than lower reference point exist in partitions segmented by the reference points despite not appearing on average.

A second contribution is to isolate effects of comparison to peers through reporting from comparison through social interaction. A body of economic research addresses comparison through social interaction (Hanushek et al. [2003], Lavy et al. [2012], Lin [2010], Lyle and Smith [2014]). Analyses of plausibly exogenous changes in peer group composition show that exposure to high performing peers leads to improvement in one's own performance, with some evidence that very high or very low performing peers carry disproportional weight in influencing

one's own performance (Lavy et al. [2012], Lazear [2001], Hoxby and Weingarth [2006]). Fundamental to the performance effects in these studies are social interactions that involve assistance from high-performers, networking, and learning through observation (Lavy et al. [2012], Lyle and Smith [2014]). Our study focuses on displaying an anonymous standard of peer performance as opposed to altering one's peer-group or mentors. We do not find a positive relationship on average between high peer-performance and own performance through reporting alone. This suggests that the interaction component of peer-performance comparison is fundamental to the positive effect of exposure to high-performing peers.

By showing the distinct effects of RPI reference point height, we also inform theory and empirical work in economic and accounting research regarding RPI-related mechanisms. First, we contribute to the literature on target setting, which notes the apparent disconnect between prescriptions that targets should be attainable infrequently and the prevalence of frequently attainable targets set in organizations (Ioannou, Serafeim, and Li [2014], Merchant and Manzoni [1989]). Targets often either explicitly contain relative performance information or allow inferring one's relative performance (Aranda et al. [2014], Bol, et al. [2010], Merchant and Manzoni [1989], Murphy [2000]). In such settings, targets plausibly assume the role of RPI reference points by serving as a standard of peer performance for comparing oneself to. A partial explanation for the prevalence of highly attainable targets, then, might be that RPI reference points have optimal performance effects at highly attainable levels. We find no performance benefit to making RPI reference points attainable less than half the time. Our findings suggest agreement between behavioral responses to RPI reference points and the prevalence of attainable targets found in practice.

Second, we show performance implications of supervisors' use of discretion to set targets. For instance, supervisors set lower RPI-based targets to mitigate fairness concerns and to avoid confrontation costs with higher-level managers (Bol et al. [2010]). Information on the initial relative performance of these groups, paired with the results of the current study on RPI reference point height, indicate that a lower reference point is likely to improve performance of individuals in the bottom half of the distribution while a higher reference point is better able to do this in the top-half of the distribution. Our results also suggest that a more optimal approach would be to customize the reference point to the individual depending on his or her initial performance. When we show a student the more effective of the two reference points depending on his or her initial performance, this baseline average 1.98 (17.56) for Δ Grade (Δ Activity Level) rises to 3.5 (24.56). By contrast, showing students the median regardless of their initial performance produces a Δ Grade of 3.04 (Δ Activity Level of 22.81). The customized approach yields a 43% (33%) larger effect.

Third, our results provide insight into motivating performance among partitions of performance that are problem areas for tournaments. The literature on tournaments and rank-based pay shows that those who are performing very well or poorly compared to a rewarded relative performance mark do worse when they notice the distance (Asch [1990], Hannan et al. [2008], Casas-Arce and Martinez-Jerez [2009]). Our findings show that displaying the median RPI reference point motivates performance among below-median performers. We also show that, when performance is outcome-based, displaying the top-quartile reference point motivates performance among top-quartile performers. We thus offer a nonfinancial performance tool—the selection of RPI reference point height—for motivating performance among groups that a

tournament would tend to leave discouraged (if a low performer) or complacent (if a high performer).

Fourth, our study focuses on behavioral responses to RPI reference points that are prevalent in employee and executive evaluations and compensation decisions. 29% of corporations in a Corporate Executive Board survey reported using forced-curve employee rankings for performance management (McGregor [2013]). Some systems include peer quartiles as RPI reference points, two of which we test behavioral responses to (Grote [2005]). As mentioned, our results suggest that behavioral responses to viewing comparison to the median (top quartile) will be most positive among low (high) performers. If managers have to choose one or the other, they could pick the reference point that motivates the group they feel is most critical to have performing well. Our results suggest that customizing the reference point based on initial performance is preferable.

At the executive level, financial statements list peer-group composition along with executive pay relative to target percentiles of the peer group (Bebchuk and Fried, [2004], Bizjak et al. [2011]). The SEC has proposed requiring companies to disclose both the percentile of the CEO's pay and a standardized measure of the company's performance relative to the compensation peer group (Securities and Exchange Commission [2015]). While the disproportional prevalence of earning and executive pay targets above the peer median has drawn widespread criticism (Bizjak et al. [2011]), we find behavioral responses that suggest relative performance targets set above the peer top-quartile might elicit performance improvement for individuals in the top half of the distribution. Future research can weigh this dynamic along with financial and career concerns in assessing the value of high targets for performance and compensation.

Finally, we contribute to accounting literature on the format of performance information reports. Studies show that order, categorization, visibility, and other display characteristics of information included in performance reports influence decisions made in stock trading and in employee evaluation (Bloomfield et al. [2006], DeBusk, Brown, and Killough [2003], Dilla and Steinbart [2005], Maines and McDaniel [2000]). Although performance reports are also often provided to individuals to aid in improving their own performance, little evidence shows how formatting the information differently alters performance effects (Yigitbasioglu and Velcu, [2012]). Our results suggest a performance benefit of customizing reference point height to an individual based on initial performance.

3.6 CONCLUSION

Our study provides some of the first evidence of the effect of providing alternatively high reference points within RPI. We also show how the effect depends on an individual's initial performance relative to the two alternatives. Further, we address the moderating role of performance-measure type by testing a process-based and an outcome-based performance measure.

We find that the effect of providing a relatively high reference point in RPI depends on one's initial performance relative to the alternatives. We test the peer top-quartile and the peer median. The effect of providing the higher rather than the lower is concave in initial performance. The effect is negative among below-median performers. The effect is positive among above average performers. In the case of an outcome-based performance measure, it is also positive for those in the top-quartile of performance. Collectively, our findings inform the selection of a reference point to drive performance in the desired partition of initial performance.

The findings also suggest that, when reports are private as in our setting, customizing the reference point based on an individual's initial performance is preferable.

Managers and government regulators can incorporate these results when selecting RPI reference points to yield desired behavior. RPI reference points are playing a growing role in settings including retail, education, energy consumption and taxpaying. The results also reveal dynamics of social comparison that help in identifying their optimal application within oft-studied systems for measuring, managing, and reporting performance.

CHAPTER 4

CONCLUSION

The implications of this dissertation are two-fold. First public and private performance reporting drive performance and can be fine-tuned to deliver the strongest performance effect. Second, in the act of reporting performance to the performers or to the general public, the nature of the reported measure is altered. The finding is akin to Heisenberg's Uncertainty Principle. In the same way that measuring a particle in the field of quantum physics requires affecting its state, reporting performance in management settings requires affecting its state. This requirement can be advantageous, in the form of performance improvement. However, it also changes the nature of the measure so that a one-unit increase means something different for the organization than it did in the absence of measurement.

In particular, the results from Chapter 2 show that publicly disclosing physician ratings drives improvement by the ratings and by undisclosed measures of quality. At the same time, the ratings become attached to their prior values. These effects are moderated by increased web traffic to the disclosed ratings, which both drive unreported performance and strengthen the stickiness of ratings at their prior values. The results from Chapter 3 show that privately disclosing students' relative performance drives performance to a differential degree depending on the reference point for comparison. The returns to providing a higher reference point are concave in an individual's initial performance relative to the two tested alternatives: median and top-quartile performance. The study also shows some evidence that the reported performance measure, though, becomes a weaker indication of improvement by unreported measures.

References

ABELER, J., A. FALK; L. GOETTE, and D. HUFFMAN. 'Reference Points and Effort Provision.' The American Economic Review 101(2) (2011): 470–492.

ALLCOTT, H. 'Social Norms and Energy Conservation.' Journal of Public Economics 95(9–10) (2011): 1082–1095.

ALLCOTT, H. and T. ROGERS. 'The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation.' American Economic Review 104(10) (2014): 3003-3037.

ANDEL, C., S. L. DAVIDOW, M. HOLLANDER, and D. A. MORENO. "The Economics of Health Care Quality and Medical Errors." Journal of Health Care Finance 39 (2012).

ARANDA, C., J. ARELLANO, and A. DAVILA. 'Ratcheting and the Role of Relative Target Setting.' The Accounting Review 89(4) (2014): 1197-1226.

ASCH, B. J. 'Do Incentives Matter? The Case of Navy Recruiters.' Industrial & Labor Relations Review 43(3) (1990): 89S–106S.

ATKINSON, J. W. 'Motivational Determinants of Risk-Taking Behavior.' Psychological Review 64(6p1) (1957).

AZMAT, G. and N. IRIBERRI. 'The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students.' Journal of Public Economics 94(7) (2010): 435–452.

BADDELEY, M. "Herding, Social Influence and Economic Decision-Making: Socio-Psychological and Neuroscientific Analyses." Philosophical Transactions of the Royal Society B: Biological Sciences 365 (2010): 281–290.

BALASUBRAMANIAN, S. K., I. MATHUR, and R. THAKUR. "The Impact of High-Quality Firm Achievements on Shareholder Value: Focus on Malcolm Baldrige and JD Power and Associates Awards." Journal of the Academy of Marketing Science 33 (2005): 413–422.

BANDIERA, O, I. BARANKAY, and I. RASUL. 'Team Incentives: Evidence from a Firm Level Field Experiment.' Journal of the European Economic Association 11(5) (2013): 1079-1114.

BANKER, R. D., and S. M. DATAR. "Sensitivity, Precision, and Linear Aggregation of Signals for Performance Evaluation." Journal of Accounting Research 27 (1989): 21–21.

BANKER, R. D., G. POTTER, and D. SRINIVASAN. "An Empirical Investigation of an Incentive Plan that Includes Nonfinancial Performance Measures." The Accounting Review 75 (2005): 65–92.

BANKER, R. D., and R. MASHRUWALA. "Simultaneity in the Relationship Between Sales Performance and Components of Customer Satisfaction. Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior 22 (2009).

BEBCHUK, L. A. and J. M. FRIED, 'Pay without Performance: Overview of the Issues.' Journal of Applied Corporate Finance 17(4) (2005): 8-23.

BENNEAR, L. S., and S. M. OLMSTEAD. "The Impacts of the "Right to Know": Information Disclosure and the Violation of Drinking Water Standards." Journal of Environmental Economics and Management 56 (2008): 117–130.

BERGER, J., C. HARBRING, and D. SLIWKA. 'Performance Appraisals and the Impact of Forced Distribution-An Experimental Investigation.' Management Science 59(1) (2013): 54-68.

BIZJAK, J., M. LEMMON, and T. NGUYEN. 'Are All CEOs Above Average? An Empirical Analysis of Compensation Peer Groups and Pay Design.' Journal of Financial Economics 100(3) (2011): 538-555.
BLANES I VIDAL, J. and M. NOSSOL. 'Tournaments Without Prizes: Evidence from Personnel Records.' Management Science 57(10) (2011): 1721-1736.
BLOOMFIELD, R.; M. NELSON; and S. SMITH. 'Feedback Loops, Fair Value Accounting and Correlated Investments.' Review of Accounting Studies 11(2-3) (2006): 377-416.
BLUNDELL, R., and M. C. DIAS. "Alternative Approaches to Evaluation in Empirical Microeconomics." Journal of Human Resources 44 (2009): 565-640.
BOL, J. C., T. M. KEUNE, E. M. MATSUMURA, and J. Y. SHIN. 'Supervisor Discretion in Target Setting: An Empirical Investigation.' The Accounting Review 85(6) (2010): 1861-1886.
BOL, J. C. "The Determinants and Performance Effects of Managers' Performance Rating Biases. The Accounting Review 86 (2011): 1549–1575.
BROWN, S. H. "Managed Care and Technical Efficiency." Health Economics 12 (2003): 149–158.
BROWN, D., L. FERRIS, D. HELLER, and L. KEEPING. 'Antecedents and Consequences of the Frequency of Upward and Downward Social Comparison at Work.' Organizational Behavior and Human Decision Processes 102 (2007): 59-75.
BROWN, D. L., S. CLARKE, and J. OAKLEY. "Cardiac Surgeon Report Cards, Referral for Cardiac Surgery, and the Ethical Responsibilities of Cardiologists.: Journal of the American College of Cardiology 59 (2010), 2378–2382.
BUTTERS, J. 'Earnings Insight S&P 500.' Earnings Insight. Factset. (2015)
CASAS-ARCE, P. and F. A. MARTINEZ-JEREZ. 'Relative Performance Compensation, Contests, and Dynamic Incentives.' Management Science 55(8) (2009): 1306–1320.
CAMPBELL, D., M. J. EPSTEIN, and F. A. MARTINEZ-JEREZ. "The Learning Effects of Monitoring." The Accounting Review 86 (2011): 1909–1934.
CARD, D., and A. KRUEGER. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." The American Economic Review 84 (1994): 772-793.
CHANDRA A., J. GRUBER, and R. McKnight. "Patient Cost-Sharing and Hospitalization Offsets in the Elderly. The American Economic Review 100 (2010): 193–213.
CHATTERJI, A. K., and M. W. TOFFEL. "How Firms Respond to Being Rated." Strategic Management Journal 31 (2010): 917–945.
CHEVALIER, J. A., and D. MAYZLIN. "The Effect of Word of Mouth on Sales: Online Book Reviews." Journal of Marketing Research 43 (2006), 345–354.
CHRISTENSEN, H. B., E. FLOYD, and M. G. MAFFETT. "The Effects of Price Transparency Regulation on Prices in the Healthcare Industry." Chicago Booth Research Paper, 2015. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2343367
CLEVELAND CLINIC. "Cleveland Clinic | Find a Doctor." 2016. Retrieved from http://my.clevelandclinic.org/staff_directory
CMS. "HHS-Operated Risk Adjustment Methodology Meeting Discussion Paper." 2016. Retrieved from https://www.cms.gov/CCIIO/Resources/Forms-Reports-and-Other-Resources/Downloads/RA-March-31-White-Paper-032416.pdf
COLUMBIA UNIVERSITY. "Courses@CU Principles of Economics." (2016). Retrieved from

http://www.coursesatcu.com/courses/1940
CROSS, S. and L. MADSON. 'Models of the Self: Self-construals and Gender.' Psychological Bulletin 12 (1997): 5-37.
CUTLER, D., and L. S. DAFNY. "Designing Transparency Systems for Medical Care Prices." New England Journal of Medicine 364 (2011): 894–895.
DAFNY, L. S. "How do Hospitals Respond to Price Changes?" The American Economic Review, 95 (2005): 1525–1547.
DAFNY, L. S., and D. DRANOVE. "Do Report Cards Tell Consumers Anything They Don't Already Know? The Case of Medicare HMOs." The Rand Journal of Economics 39 (2008): 790–821.
DANIELS, C. and T. MILLER. 'After the Architect, Building Culture Calls for a Contractor.' Leadership Case Studies (2014).
DEBUSK, G. K., R. M. BROWN, and L. N. KILLOUGH. 'Components and Relative Weights in Utilization of Dashboard Measurement Systems Like the Balanced Scorecard.' The British Accounting Review 35(3) (2003): 215-231.
DILLA, W. N. and P. J. STEINBART. 'The Effects of Alternative Supplementary Display Formats on Balanced Scorecard Judgments.' International Journal of Accounting Information Systems 6(3) (2005): 159–176.
DOLAN, P., M. HALLSWORTH, D. HALPERN, D. KING, R. METCALFE, and I. VLAEV. 'Influencing Behaviour: The Mindspace Way.' Journal of Economic Psychology 33(1) (2012): 264–277.
DOYLE, J. J., Jr. "Returns to Local-Area Health Care Spending: Evidence from Health Shocks to Patients Far from Home." American Economic Journal: Applied Economics 3 (2011): 221–243.
DRANOVE, D., D. KESSLER, M. MCCLELLAN, and M. SATTERTHWAITE. "Is More Information Better? The Effects of "Report Cards" on Health Care Providers." The Journal of Political Economy 111 (2003): 555–588.
DUFLO, E. Empirical Methods. Massachusetts Institute of Technology. (2002). Retrieved from http://dspace.mit.edu/bitstream/handle/1721.1/49516/14-771Fall-2002/NR/rdonlyres/Economics/14-771Development-Economics--Microeconomic-Issues-and-Policy-ModelsFall2002/2494CA2C-D025-40A6-B167-F8A5662520DB/0/emp_handout.pdf
DYNARSKI, S. "Hope for Whom? Financial Aid for the Middle Class and Its Impact on College Attendance." National Tax Journal (2000): 629-661.
EAGLY, A. H. 'Sex Differences in Influenceability.' Psychological Bulletin 85 (1978): 86-116.
EPSTEIN, A. J. "Do Cardiac Surgery Report Cards Reduce Mortality? Assessing the Evidence. Medical Care Research and Review 63 (2006): 403–426.
EREZ, M. and I. ZIDON. 'Effects of Goal Acceptance on the Relationship of Goal Setting and Task Performance.' Journal of Applied Psychology 69 (1984): 69-78.
EREZ, M., P. C. EARLEY, and C. L. HULIN. 'The Impact of Participation on Goal Acceptance and Performance: A Two Step Model.' Academy of Management Journal 28 (1985): 50-66.
ERYARSOY, E., and S. PIRAMUTHU. "Experimental Rating of Sequential Bias in Online Customer Reviews." Information and Management, 51 (2014): 964–971.
EVANS, I., Y. HWANG, and N. J. NAGARAJAN. "Management Control and Hospital Cost Reduction: Additional Evidence." Journal of Accounting and Public Policy 20 (2001): 73–88.

FARBER, H. S. 'Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers.' American Economic Review (2008) 98(3): 1069-1082.
FESTINGER, L. 'A Theory of Social Comparison Processes.' Human relations 7(2) (1954): 117-140.
FISHER, J., S. PEFFER, and G. SPRINKLE. 'Budget-Based Contracts, Budget Levels, and Group Performance.' Journal of Management Accounting Research 15(1) (2003): 51-74.
FLAHERTY, C. "Evaluating Evaluations." (2014). Retrieved from https://www.insidehighered.com/news/2014/05/20/study-suggests-research-plays-bigger-role-faculty-evaluations-student-evaluations
FURNHAM, A., and H. C. BOO. "A Literature Review of the Anchoring Effect." The Journal of Socio-Economics 40 (2011): 35–42.
GARCIA, S. and A. TOR. 'Rankings, Standards, and Competition: Task Versus Scale Comparisons.' Organizational Behavior and Human Decision Processes 102 (2007): 95-108.
GIBBONS, R and J. ROBERTS. The Handbook of Organizational Economics. Princeton University Press, 2012.
GLOVER, L. "Are Online Physician Ratings any Good?" (2014). Retrieved from http://health.usnews.com/health-news/patient-advice/articles/2014/12/19/are-online-physician-ratings-any-good
GORMAN, A. 'How One Hospital Reduced Unnecessary C-Sections.' The Atlantic, 2015. Available at: http://www.theatlantic.com/health/archive/2015/05/how-one-hospital-reduced-unnecessary-c-sections/392924/
GRAHAM, M. "Regulation by Shaming." (2000). Retrieved from http://www.theatlantic.com/magazine/archive/2000/04/regulation-by-shaming/378126/
GROTE, R. C. Forced Curve: Making Performance Management Work. Harvard Business School Press, 2005.
HALLSWORTH, M., J. A. LIST, R. D. METCALFE, and I. VLAEV. 'The Behavioralist as a Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance.' NBER Working Paper Series, 2014. Available at: http://www.nber.org/papers/w20007.pdf
HANAUER, D. A., K. ZHENG, D. C. SINGER, A. GEBREMARIAM, and M. M. DAVIS. "Public Awareness, Perception, and Use of Online Physician Rating Sites." JAMA 311 (2014): 734.
HANNAN, E. L., M. S. SARRAZIN, D. DORAN, and G. ROSENTHAL. "Provider Profiling and Quality Improvement Efforts in Coronary Artery Bypass Graft Surgery: The Effect on Short-Term Mortality Among Medicare Beneficiaries." Medical Care 41 (2003).
HANNAN, L., R. KRISHNAN, and A. NEWMAN. 'The Effects of Disseminating Relative Performance Feedback in Tournament and Individual Performance Compensation Plans.' The Accounting Review 83(4) (2008): 893–913.
HANNAN, L., G. MCPHEE, A. NEWMAN, and I. TAFKOV. 'The Effect of Relative Performance Information on Performance and Effort Allocation in a Multi-Task Environment.' The Accounting Review 88(2) (2013): 553-575.
HANUSHEK, E. A.; J.F. KAIN; J. M. MARKMAN; and S. G. RIVKIN. 'Does Peer Ability Affect Student Achievement? Empirical Analysis of Social Interactions.' Journal of Applied Econometrics 18(5) (2003): 527-544.
HARPER, F. M., J. KONSTAN, Y. CHEN, and S. X. LI. 'Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens.' The American Economic Review 100(4) (2010): 1358–1398.

HEATH, C; R. LARRICK; and G. WU. 'Goals as reference points." Cognitive psychology 38(1) (1999): 79-109.
HÖLMSTROM, B. "Moral Hazard and Observability." The Bell Journal of Economics (1979): 74–91.
HU, N., L. LIU, and J. J. Zhang. "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects." Information Technology and Management 9 (2008): 201–214.
IOANNOU, I; S. LI; and G. SERAFEIM. 'The Effect of Target Difficulty on Target Completion: The Case of Reducing Carbon Emissions.' Unpublished Paper (forthcoming in The Accounting Review), 2014. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2133004
ITTNER, C. D., and D. F. LARCKER, "Are Nonfinancial Measures Leading Indicators of Financial Performance?" An Analysis of Customer Satisfaction." Journal of Accounting Research 36 (1998): 1–35.
ITTNER, C. D., D. F. LARCKER, and M. W. MEYER. "Subjectivity and the Weighting of Performance Measures: Evidence from a Balanced Scorecard." The Accounting Review 78 (2003): 725–758.
JIN, G. Z., and P. LESLIE. "The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards." The Quarterly Journal of Economics, (2003): 409–451.
JOHN, P.; M. SANDERS; and J. WANG. 'The Use of Descriptive Norms in Public Administration: A Panacea for Improving Citizen Behaviours?' Unpublished paper, 2014. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2514536.
JOYNT, K. E., E. J. ORAV, and A. K. JHA. 'Thirty-day Readmission Rates for Medicare Beneficiaries by Race and Site of Care. JAMA, 305 (2011), 675–681.
KAHNEMAN D. and A. TVERSKY. 'Prospect Theory: An Analysis of Decision Under Risk." Econometrica 47(2) (1979): 263-292.
KAPLAN, R. S., and D. P. NORTON. "The Balanced Scorecard: Measures that Drive Performance." Harvard Business Review 83 (2005): 172.
KEYNES, J. M. "A Treatise on Money." (1930).
KLAR Y. and E. E. GILADI. 'No One in my Group can be Below the Group's Average: a Robust Positivity Bias in Favor of Anonymous Peers.' Journal of Personality and Social Psychology 73(5) (1997): 885-901.
KETTLE, S., M. HERNANDEZ, S. RUDA, and M. SANDERS. 'Behavioural Insights to Improve Tax Compliance: Short-Term Impacts from a Randomised Experiment in Guatemala.' (2015) CMPO Working Paper.
KOLSTAD, J. T. "Information and Quality when Motivation is Intrinsic: Evidence from Surgeon Report Cards." The American Economic Review, 103 (2003): 2875–2910.
LAMB, A., J. SMILACK, A. D. HO, and J. REICH. 'Addressing Common Challenges in Randomized Experiments in MOOCs: A Study of Encouraging Discussion in JusticeX'. Proceedings of the Second ACM Conference on Learning@Scale (2015). Available at: http://harvardx.harvard.edu/files/harvardx/files/mooc_analytic_challenges_harvardx_wp.pdf
LARRICK, R. P., K. A. BURSON, and J. B. SOLL. 'Social Comparison and Confidence: When Thinking You're Better than Average Predicts Overconfidence (and when it does not).' Organizational Behavior and Human Decision Processes 102(1) (2007): 76–94.
LAVY, V., O. SILVA, and F. WEINHARDT. 'The Good, the Bad, and the Average: Evidence on Ability Peer Effects in Schools.' Journal of Labor Economics 30(2) (2012): 367-414.

LAWLER, E. E. 'A Correlational-Causal Analysis of the Relationship Between Expectancy Attitudes and Job Performance.' Journal of Applied Psychology 52(6p1) (1968).
LAWLER, E. E. and J. L. SUTTLE. 'Expectancy Theory and Job Behavior.' Organization Behavior and Human Performance 9 (1973): 482-503.
LAZEAR, E. 'Educational Production.' Quarterly Journal of Economics 116(3) (2001): 777-803.
LAZEAR, E. and S. ROSEN. 'Rank-order Tournaments as Optimum Labor Contracts.' Journal of Political Economy 89(5) (1981): 841-864.
LEUZ, C., and P. WYSOCKI. "The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research." Journal of Accounting Research 54 (2016): 525–622.
LIN, XU. 'Identifying Peer Effects in Student Academic Achievement by Spatial Autoregressive Models with Group Unobservables.' Journal of Labor Economics 28(4) (2010): 825-860
LOCKE, E. A. and J. BRYAN. 'The Directing Function of Goals in Task Performance.' Organizational Behavior and Human Performance 4 (1969): 35-42.
LOCKE, E. A. and G. P. LATHAM. "Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-year Odyssey." American Psychologist 57(9) (2002).
LOCKE, E. A., S. J. MOTOWIDLO, and P. BOBKO, 'Using Self-Efficacy Theory to Resolve the Conflict Between Goal-Setting and Expectancy Theory in Organizational Behavior and Industrial/Organizational Psychology.' Journal of Social and Clinical Psychology 4 (1986): 328-338.
LYLE, D. S., and J. Z. SMITH. 'The Effect of High-Performing Mentors on Junior Officer Promotion in the US Army' Journal of Labor Economics 32(2) (2014): 229–258.
LU, F. S. "Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes." Journal of Economics and Management Strategy 21 (2012): 673–705.
LUCA, M. "Reviews, Reputation, and Revenue: the Case of Yelp.Com." SSRN Electronic Journal (2016). Available at: http://doi.org/10.2139/ssrn.1928601
LYU, H., E. C. WICK, M. HOUSMAN, J. A. FREISCHLAG, and M. A. MAKARY. "Patient Satisfaction as a Possible Indicator of Quality Surgical Care." JAMA Surgery 148 (2013): 362.
MAINES, L. A. and L. S. MCDANIEL. 'Effects of Comprehensive-Income Characteristics on Nonprofessional Investors' Judgments: The Role of Financial-Statement Presentation Format.' The Accounting Review 75(2) (2000): 179-207.
MARRIOTT. "Reviews." 2016. Retrieved from http://www.marriott.com/hotels/hotel-reviews/chijw-jw-marriott-chicago/
MCGREGOR. 'For Whom the Bell Curve Tolls.' The Washington Post, 2013. Available at: https://www.washingtonpost.com/news/on-leadership/wp/2013/11/20/for-whom-the-bell-curve-tolls/
MERCHANT, K. A. and J. F. MANZONI.' The Achievability Of Budget Targets In Profit Centers: A Field Study.' The Accounting Review 64(3) (1989).
MERLINO, J. I., and A. RAMAN. "Health Care's Service Fanatics." Harvard Business Review 91 (2013): 108-16.
MOERS, F. "Discretion and Bias in Performance Rating: the Impact of Diversity and Subjectivity." Accounting, Organizations and Society 30 (2005): 67–80.
MUCHNIK, L., S. ARAL, and S. J. TAYLOR. "Social Influence Bias: A Randomized Experiment." Science 341 (2013): 647–651.
MURPHY, K. J. 'Performance Standards in Incentive Contracts.' Journal of Accounting and

Economics 30(3) (2000): 245-278.
MURPHY, K. J., and T. SANDINO. "Executive Pay and “Independent” Compensation Consultants." Journal of Accounting and Economics, 49 (2010): 247-262.
MURTHY, U. 'The Effect of Relative Performance Information under Different Incentive Schemes on Performance in a Production Task.' AAA 2011 Management Accounting Section (MAS) Meeting Paper, 2010. Available at: http://ssrn.com/abstract=1632663
NAGAR, V., and M. V. RAJAN. "Measuring Customer Relationships: The Case of the Retail Banking Industry." Management Science 51 (2005): 904-919.
NEWMAN, B. M., and P. R. NEWMAN. "Development Through Life: A Psychosocial Approach. Cengage Learning, 2014.
PARKER, C., and V. L. NIELSEN. "Explaining Compliance: Business Responses to Regulation." Edward Elgar Publishing, 2011.
PERKINS, H. W., M. P. HAINES, and R. RICE. 'Misperceiving the College Drinking Norm and Related Problems: A Nationwide Study of Exposure to Information, Perceived Norms, and Student Alcohol Misuse.' Journal of Studies on Alcohol and Drugs 66(4) (2005): 470-478.
PETERSEN, M. A. Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. Review of Financial Studies 22 (2009): 435–480.
PETERSON, E. D., E. R. DELONG, J. G. JOLLIS, L. H. MUHLBAIER, D. B. MARK, "The Effects of New York’s Bypass Surgery Provider Profiling on Access to Care and Patient Outcomes in the Elderly." Journal of the American College of Cardiology 32 (1998): 993–999.
PRENDERGAST, C., and R. TOPEL. "Discretion and Bias in Performance Rating." European Economic Review 37 (2003): 355–365.
RENO, R. R., R. B. CIALDINI, and C. A. KALLGREN. 'The Transsituational Influence of Social Norms.' Journal of Personality and Social Psychology 64(1) (1993): 104–112.
ROGERS, T. and A. FELLER. 'The Threat of Excellence: Exposure to Peers' Exemplary Work Undermines Performance and Success.' Presented at the Society for Judgement and Decision Making 2015 36th Annual Conference (2015).
RYAN, A. M., B. K. NALLAMOTHU, and J. B. DIMICK. "Medicare’s Public Reporting Initiative on Hospital Quality had Modest or no Impact on Mortality from Three Key Conditions." Health Affairs, 31 (2012): 585–592.
SECURITIES AND EXCHANGE COMMISSION, 'SEC Proposes Rules to Require Companies to Disclose the Relationship Between Executive Pay and a Company's Financial Performance.' Press Release. SEC. 2015.
SEDIKIDES, C, L. GAERTNER, and Y. TOGUCHI. 'Pancultural self-enhancement.' Journal of Personality and Social Psychology 84(1) (2003): 60–79.
SCHULTZ, P. W.; J. M. NOLAN; R. B. CIALDINI; N. J. GOLDSTEIN; and V. GRISKEVICIUS. 'The Constructive, Destructive, and Reconstructive Power of Social Norms.' Psychological Science 18(5) (2007): 429–434.
SIKORA, R. T., and K. CHUAHAN. "Estimating Sequential Bias in Online Reviews: A Kalman Filtering Approach." Knowledge-Based Systems, 27 (2012): 314–321.
SMITH, R. "Assimilative and Contrastive Emotional Reactions to Upward and Downward Social Comparisons." In Handbook of Social Comparison, ed. J. Suls and L. Wheeler, 2000.
SONG, H., A. TUCKER, K. MURRELL, and D. VINSON, 'Public Relative Performance Feedback in Complex Service Systems: Improving Productivity through the Adoption of

Best Practices.' Unpublished paper, 2015. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2673829
SHUKLA, M. "Long-Term Impact of Coronary Artery Bypass Graft Surgery (CABG) Report Cards on CABG Mortality and Provider Market Share and Volume." The George Washington University, 2013. Ed. A. Dor. Retrieved from http://media.proquest.com/media/pq/classic/doc/3085535221/fmt/ai/rep/NPDF?_s=TO9PEYnJkbXRA%2F8fETo9otvBzqw%3D
SVENSON, O. 'Are we all Less Risky and More Skillful than our Fellow Drivers?' Acta Psychologica, 47(2) (1981): 143–148.
STANFORD HEALTH CARE. "Find a Doctor | Stanford Health Care." 2016. Retrieved from https://stanfordhealthcare.org/search-results.doctors.html
STARWOOD. "Guest Ratings and Reviews | Sheraton New York Times Square Hotel." 2016. Retrieved from http://www.starwoodhotels.com/sheraton/property/reviews/index.html?propertyID=421
SUNDARARAJAN, V., T. HENDERSON, C. PERRY, A MUGGIVAN, H. QUAN, and W. A. GHALI. "New ICD-10 Version of the Charlson Comorbidity Index Predicted In-Hospital Mortality. Journal of Clinical Epidemiology, 57 (2004): 1288–1294.
TAFKOV, I. D. "Private and Public Relative Performance Information Under Different Compensation Contracts." The Accounting Review 88 (2013): 327–350.
TEXAS TECH UNIVERSITY. "Student Ratings of Courses and Instructors." (2016). Retrieved from http://www.ttu.edu/courseinfo/evals/
TVERSKY, A., and D. KAHNEMAN. "Judgment Under Uncertainty: Heuristics and Biases." In Utility, Probability, and Human Decision Making, eds. D. Wendt and C. Vlek (1995).
UBEL, P. "Paying for Patient Satisfaction Harms Hospitals that Care for Poor People." (2015). Retrieved from http://www.forbes.com/sites/peterubel/2015/10/16/paying-for-patient-satisfaction-harms-hospitals-that-care-for-poor-people/#39a7355b62f7
VAN DER STEDE, W. "Management Accounting: From Where, Where Now, Where To?" Journal of Management Accounting Research 17 (2015). 171-176.
VANEK SMITH, S. 'I, Waiter' NPR. (2015) Available at http://www.npr.org/templates/transcript/transcript.php?storyID=407086723
VIVES, X. "Information and Learning in Markets: the Impact of Market Microstructure." Princeton University Press, 2010
VROOM, V. H. Work and Motivation. Wiley, 1964.
WEIL, D., A. FUNG, M. GRAHAM, and E. FAGOTTO. "The Effectiveness of Regulatory Disclosure Policies." Journal of Policy Analysis and Management 25 (2006), 155–181.
WINDSCHITL, P. D., J. KRUGER, and E. N. SIMMS. 'The Influence of Egocentrism and Focalism on People's Optimism in Competitions: When What Affects Us Equally Affects Me More.' Journal of Personality and Social Psychology 85(3) (2003): 389–408.
XIAO, Y. and R. LUCKING. 'The Impact of Two Types of Peer Assessment on Stuents' Performance and Satisfaction within a Wiki Environment'. Internet and Higher Education 11(3-4) (2008): 186-193.
YIGITBASIOGLU, O., and O. VELCU. 'A Review of Dashboards in Performance Management: Implications for Design and Research.' International Journal of Accounting Information Systems 13(1) (2012): 41–59.

Appendix A: Patient Satisfaction Survey Questions Used in Disclosure

Physician (referred to as "care provider" in the survey questions)

1) Friendliness/courtesy of the care provider
2) Explanations the care provider gave you about your problem or condition
3) Concern the care provider showed for your questions or worries
4) Care provider's efforts to include you in decisions about your treatment
5) Degree to which care provider talked with you using words you could understand
6) Amount of time the care provider spent with you
7) Your confidence in this care provider
8) Likelihood of your recommending this care provider to others
9) Length of wait time at clinic

Appendix B: Example Patient Comments

Each patient satisfaction survey contains a text box under the Care Provider section of a survey with the prompt: "Comments (Describe Good or Bad Experience)" that patients can choose to fill in. Comments regarding physicians were posted in their entirety on the official online profiles of included in disclosure, except when administrators, who were not physicians, screened comments that contained slander or personally identifiable information about the patient. The following are example comments:

#	**Selected Comment**
1	The only complaint I had was that he didn't tell me how many precancer spots I had on my face so I wasn't prepared for as many as there were. I wasn't quite mentally prepared for getting sprayed with the liquid nitrogen that many times.
2	Excellent service all the way around.
3	Dr. Salari is the best physician there in my opinion. I haven't really seen any others but I trust him to give me an honest opinion and he shows he cares. Apparently everyone likes him because at times he is hard to get an appointment with but that I guess is a good thing
4	I would have liked to have heard a bit about the down-side of "prednisone" – such as it can sometimes cause mood swings or depression. I wasn't prepared for "feeling so down".
5	The symptoms I had were frightening and the physician was very good at explaining what was going on and alleviating my fears.

Appendix C: Variable Definitions

Chapter 2 Variables

Dependent Variables	**Description**
Rating	The average of the nine component ratings regarding a physician in a Press Ganey survey returned following a patient visit. Each component rating is on a Likert scale of one to five, with five as the most favorable rating.
Quality deduction	An indicator variable equal to one if the visit resulted in a readmission to the emergency department within 30 days of discharge, a hospital-acquired condition, or both.
Absolute difference	The absolute distance of an individual physician rating from the consensus rating as calculated for disclosure. For physicians included in the December 2012 posting, the consensus was calculated and disclosed, and for those excluded I retrospectively calculate the consensus. For all physicians, I retrospectively calculate the consensus as would have been disclosed in December 2011 had the disclosure occurred then. Absolute difference is measured relative to the December 2011 consensus prior to disclosure, and to the December 2012 consensus following disclosure and prior to the July 2013 rating posting/update.
Treatment Variable	**Description**
Disclosed	An indicator for the time period following which a physician's ratings were disclosed, if ever.
Placebo disclosed	For the tests of Models 1 and 2, regarding performance effects, this is an indicator for the time period beginning one year prior to the first time at which a physician's rating were disclosed, if ever. For the tests of Model 3, regarding consensus bias, this is an indicator for the time period beginning seven months prior to the December 2012 disclosure for physicians included in the disclosure.
Partitioning Variable	**Description**
Web traffic	The number of page views of the physician's official online profile in the calendar month prior to the observed visit.
Control Variables	**Description**
Age	Patient age at the time of the visit, with ages above 89 treated as 90. For the tests of Models 1 and 3, related to physician ratings and bias toward ratings, ages are included as dummies using the psychometric categories of Newman and Newman (2014): 0-11, 11-17, 18-24, 25-34, 45-59, 59-74, +74. For the tests of Model 2, regarding quality, ages are included as outlined by CMS (2016) for use in risk adjustment: 0-1, 1-4, 5-9, 10-14, 15-20, 21-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60+.
Gender	An indicator variable equal to 1 if the physician is female.
Medicare or Medicaid	An indicator variable equal to 1 if Medicare or Medicaid was the primary insurance used for the visit.

Severity/complexity	A component of Medicare reimbursement formulas that accounts for the patient's case severity and the complexity of care provided.
Comorbidity	The Charlson Comorbidity Index, which takes a value of one, two, three, or six in proportion to the likelihood of mortality within one year associated with the comorbid condition. Comorbid conditions include heart disease, aids, and cancer among the 22-condition set. The conditions are recorded at the time of a procedure. Thus, for regressions on quality, the value is included as assigned to the given visit. For regressions on patient satisfaction ratings or their derivatives, which may occur before or after a procedure, the value is included as the six-month rolling window within the sample centered at the patient visit. The results are robust to narrowing the window to three months or expanding it to one year.
Charges	The dollar value of charges assigned to the visit.
First visit	An indicator variable equal to 1 if the visit is the patient's first to the physician conducting the visit.
Physician week's visit count	The total number of visits conducted by the physician conducting a visit in the same week.
English speaking	An indicator variable equal to 1 if a patient indicated in a survey response that they speak English.
Contemporary standard deviation	The standard deviation of the physician's ratings in the period in which the rating occurred relative to the December 2012 rating posting.
Consensus count	The inverse square root of the count of observations that comprise a physician's consensus rating as calculated for disclosure and used in this study in measuring absolute difference.
Rating trend	The rating trend for a given physician in the period in which the rating occurred relative to the December 2012 rating posting.
Year	A categorical variable for the calendar year in which the visit occurred: 2011, 2012, 2013, or 2014.
Period	A categorical variable for the period, segmented by disclosure events (i.e., the December 2012 and the July 2013 rating postings), in which the visit occurred: 1 for before the first posting, 2 for after the first and before the second posting, and 3 for after both postings.
Physician dummies	An indicator variable for the physician conducting the visit.
Physician Variables	**Description**
Age	The physician's age as of January 1, 2011.
Gender	An indicator variable equal to 1 if the physician is a female.
MD	An indicator variable equal to 1 if the physician holds an MD.
Years with UUHC	The number of years that UUHC has employed the physician.
Tenure track	An indicator variable equal to 1 if the physician has a tenure-track appointment.

Dependent Variables	**Description**
Activity Level	The following weighted sum, that approximately scales each type of action's historical mean to the historical mean of video views in the experiment host courses: video views + 1.5 x problem attempts + 20 x forum posts + 2.5 x other forum actions + 5 x number of days active in the course + 0.1 x total actions
Δ Activity Level	Activity Level at the experiment's end minus Activity Level at the experiment's beginning
Grade	Grade in the course
Δ Grade	Grade at the experiment's end minus Grade at the experiment's beginning
Dependent Variable Components	**Description**
Video Views	The number of times a student started watching a video
Problem Attempts	The number of times a student entered an answer to any problem
Forum Posts	The number of posts a student made in discussion forums
Other Forum Actions	The number of actions (e.g., voting for a post, responding with a comment to an original post) a student took in discussion forums
Number of Days Active in the Course	The number of calendar days on which a student accessed the course
Total Actions (component of Activity Level)	All actions in the course that are recorded electronically; these include video views, problem attempts, forum posts, and other forum actions, but are not limited to them
Independent Variables	**Description**
Control	An indicator variable equal to one if the individual is a member of the control group, which does not receive RPI displays.
Median Reference Point (RPI_M)	An indicator variable equal to one if the individual is a member of the treatment group that receives an RPI display with the peer median reference point; the peer median is the median activity level of individuals who have accessed the course
Top-quartile Reference Point (RPI_T)	An indicator variable equal to one if the individual is a member of the treatment group that receives an RPI display with the peer top-quartile reference point; the peer top-quartile is the top-quartile activity level of individuals with who have accessed the course
Relative Performance Information (RPI)	An indicator variable equal to one if either Top-quartile Reference Point = 1 or Median Reference Point = 1
Moderator Variables	**Description**
Initially Below Average	An indicator variable equal to one if the individual's

	activity level was less than or equal to the median of all individuals who had accessed the course at the experiment's start
Initially Third Quartile	An indicator variable equal to one if the individual's activity level was greater than the median and less than the top-quartile of all individuals who had accessed the course at the experiment's start
Initially Top Quartile	An indicator variable equal to one if the individual's activity level was greater than or equal to the top-quartile of all individuals who had accessed the course at the experiment's start
Gender	An indicator variable equal to one if the individual indicated their gender in the course registration process and chose Female, and equal to zero if the individual indicated their gender in the course registration process and chose Male.
Developed Country	An indicator variable equal to one if the individual indicated their country of residence and the country is of UN Developed Nation status, and equal to zero if the individual indicated their country of residence and the country is of UN Developing Nation Status
Level of Education	An indicator variable equal to one if the individual indicated their level of education and has a bachelor's degree or higher degree.
Age	The age, if any, that an individual indicated during registration, truncated at 5 and 100.
Descriptive Variables	**Description**
Familiarity With Subject	Response to survey question, "How familiar are you with [course]?" 0 = Not at all Familiar; 1 = Slightly Familiar; 2 = Somewhat Familiar; 3 = Very Familiar; 4 = Extremely Familiar.
Commitment to Complete Course	Response to survey question, "People register for HarvardX courses for different reasons. Which of the following best describes you?" 1 = Here to brows the materials, but not planning on completing any course activities (watching videos, reading text, answering problems, etc.); 2 = Planning on completing some course activities, but not planning on earning a certificate; 3 = Planning on completing enough course activities to earn a certificate.
Number of Online Courses Previously Enrolled-In	Response to survey question, "How many online courses have you *registered* for in the past?"
Number of Online Courses Previously Completed	Response to survey question, "How many online courses have you *completed* in the past?"

Appendix D: Relative Performance Information Displays, Emails and Surveys

1. Example Email with Link to RPI Display

From: HarvardX <noreply@qemailserver.com> on behalf of HarvardX <noreply@qemailserver.com>
Date: Friday, March 25, 2016 at 9:02 AM
To: Jane Doe <username@domain.com>
Subject: Your Personalized Activity Graph for MCB80.1x

Dear Learner,

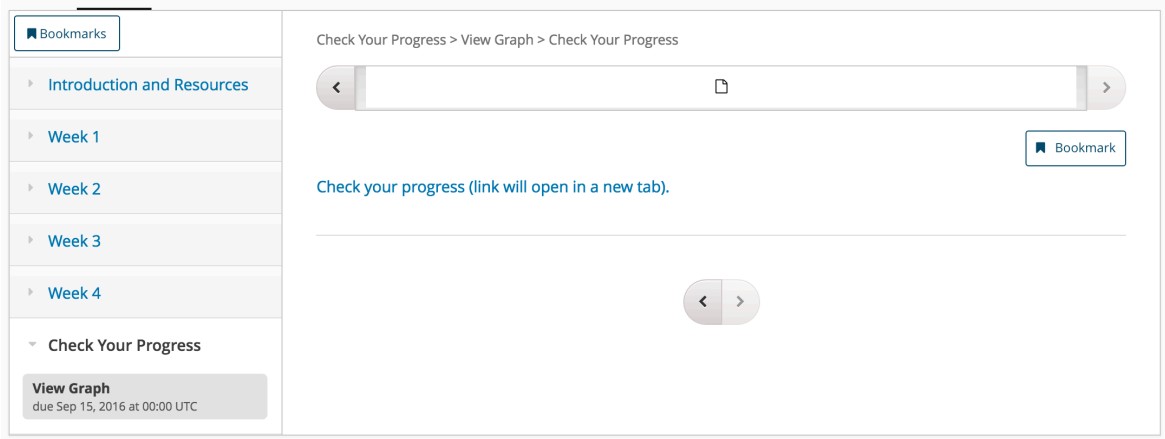
We've built a personalized comparison graph of your activity in MCB80.1x. The graph will self-update daily as you do more in the class. Your most up-to-date graph will always be at this link: see your graph.

Sincerely,
The Team at HarvardX

The ability to send you these weekly emails aids researchers, who are very grateful if you are willing to receive them. The emails will stop arriving when the course ends. If you prefer to stop recieving them immediately, follow this link: Click here to unsubscribe.

In the supplemental experiment, the "Grade" replaces any reference to "Activity"

2. Example In-Course Link to RPI Display

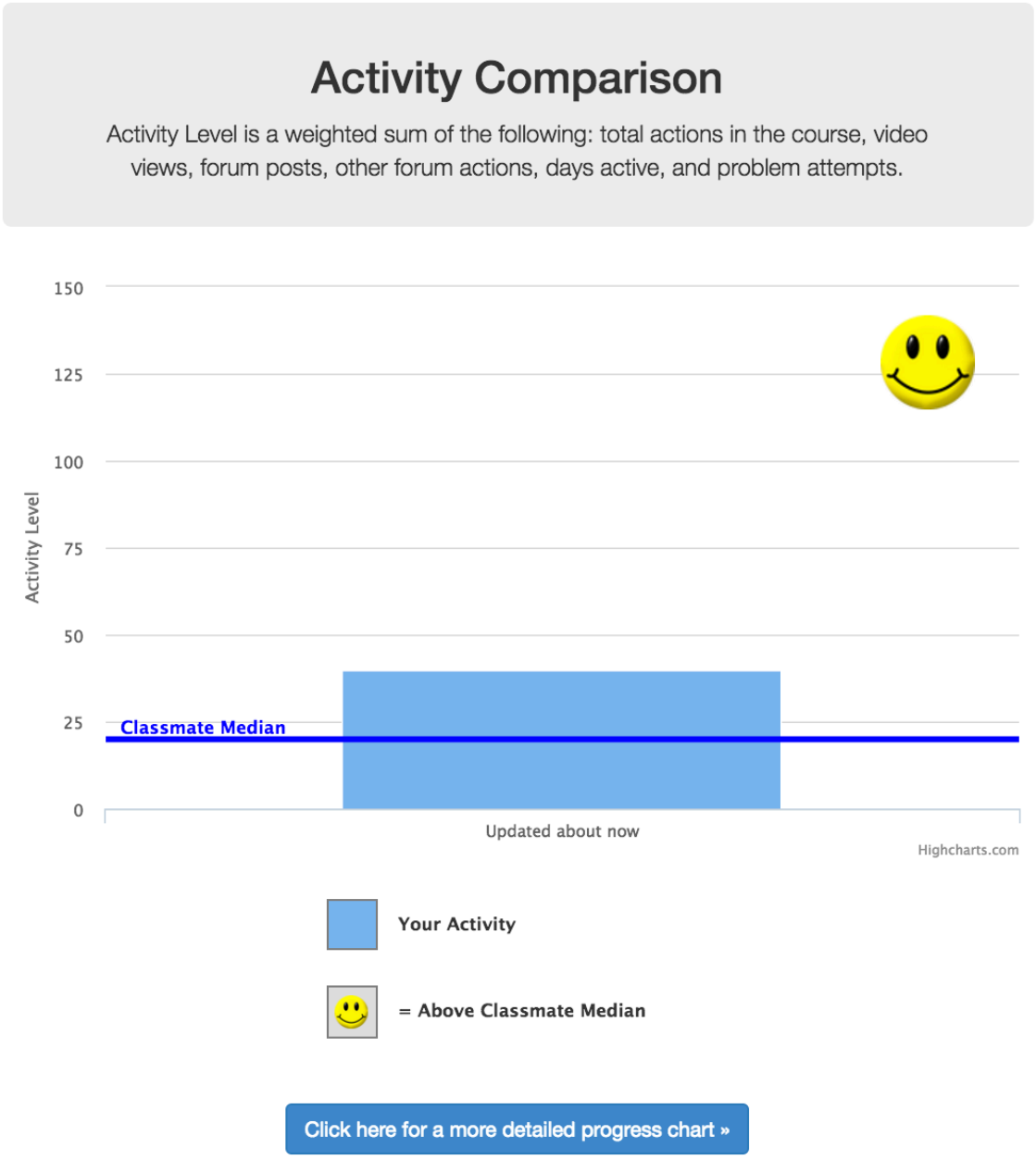


The link titled, "Check your progress (link will open in a new tab)" takes control-group students to the standard course progress chart for HarvardX courses. The same link takes treatment group students directly to the proposed experiment's RPI display that is customized to their activity in the course. The RPI display webpage has a link at the

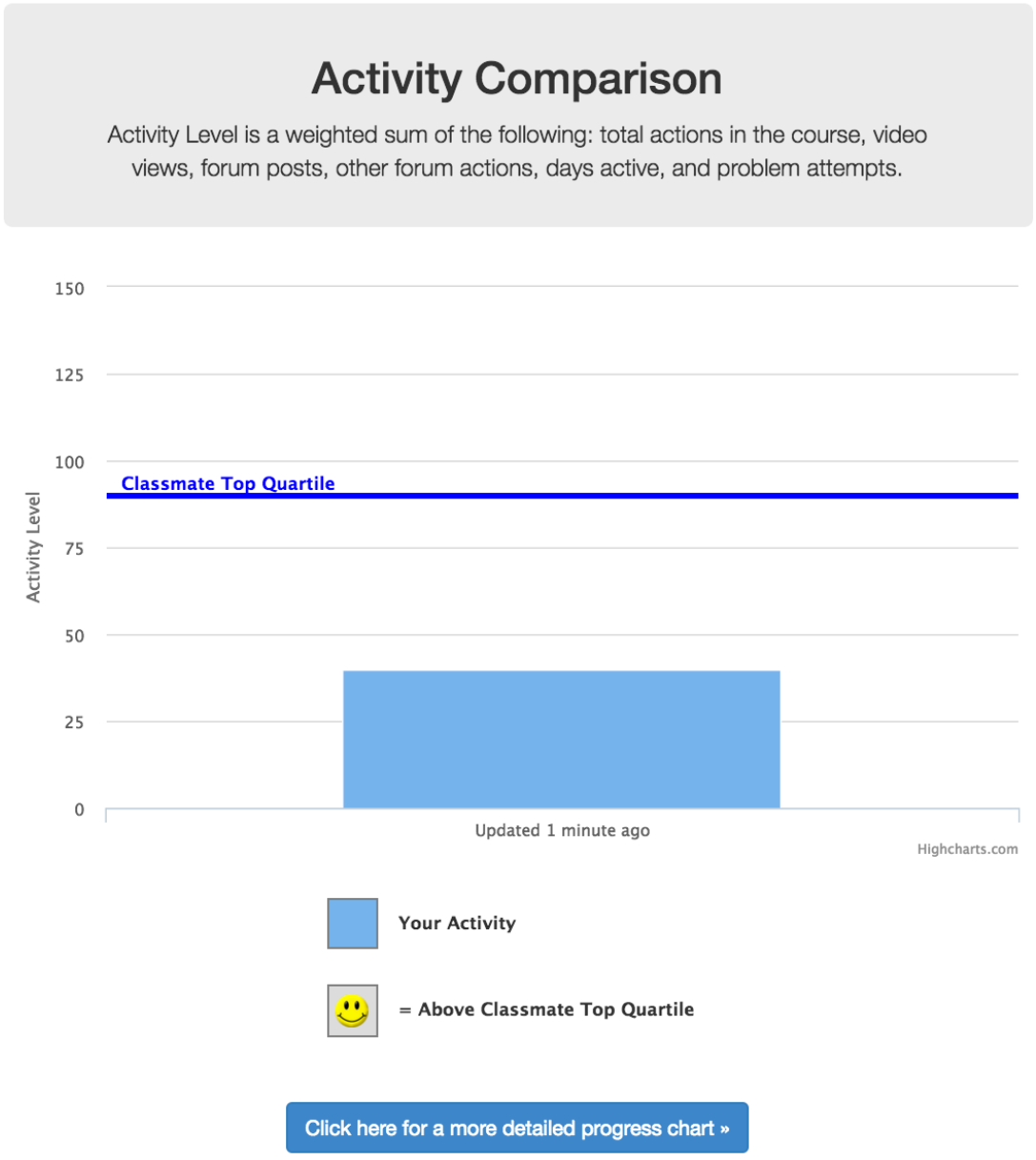
bottom titled, “Click here for a more detailed progress chart”, which takes treatment-group students to the standard HarvardX course progress chart.

3. Activity Level Displays

Peer-Median Reference Point

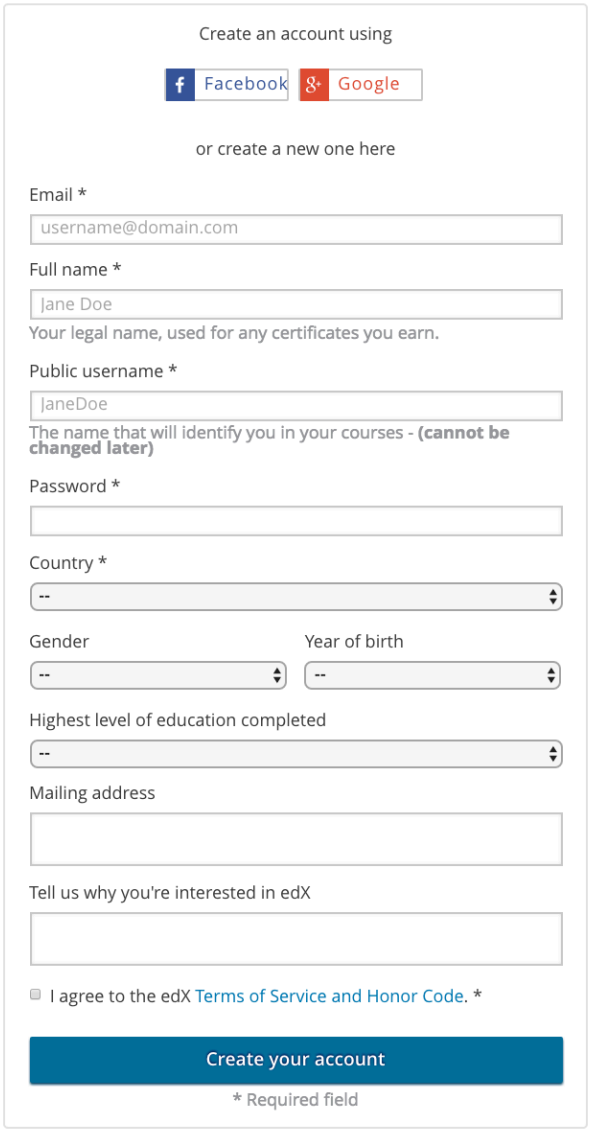


Peer-Top-Quartile Reference Point



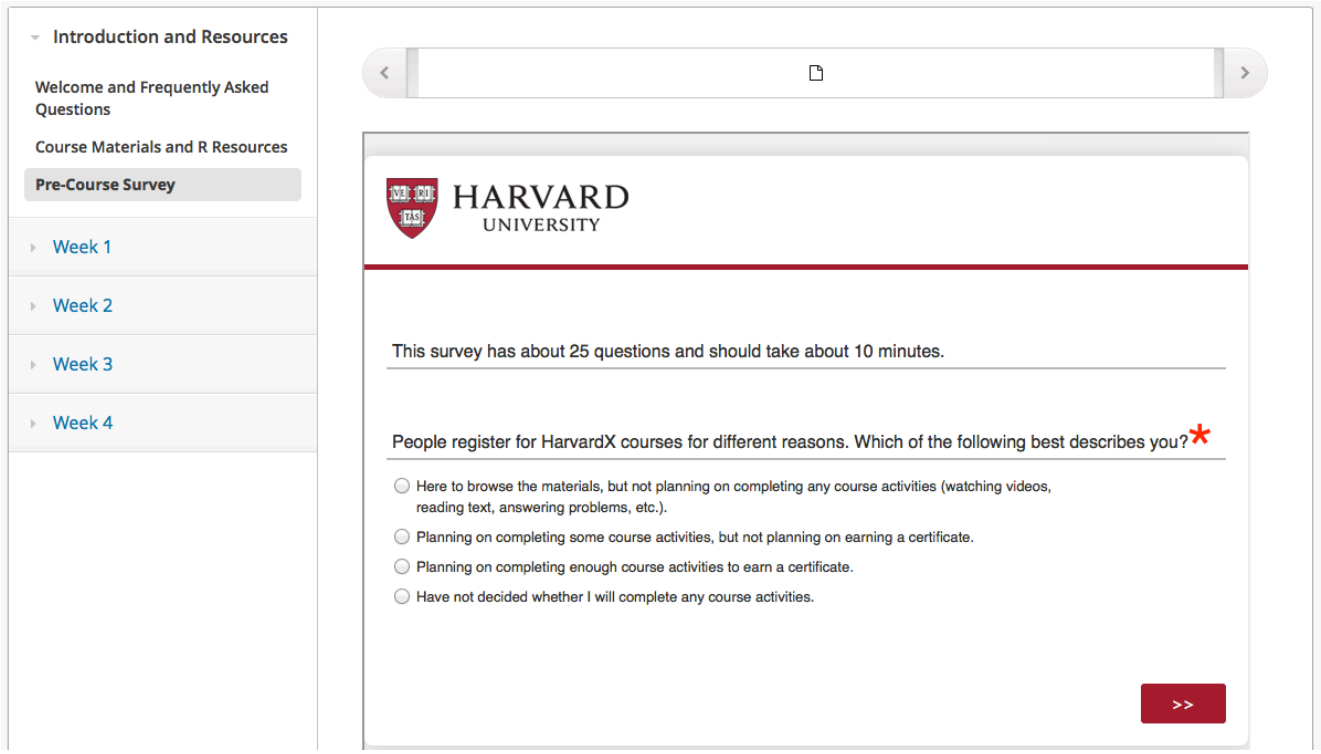
In the main experiment, the graphs dynamically scale to show levels of activity above 150, and start with a default height of 150. The "Click here for a more detailed progress chart" link takes treatment group students to the default HarvardX course progress chart. In the supplemental experiment, the "Grade" replaces any reference to activity or activity level. In that experiment, the graph has a fixed scale from 0-100.

4. Registration Form (Required of all HarvardX Students)

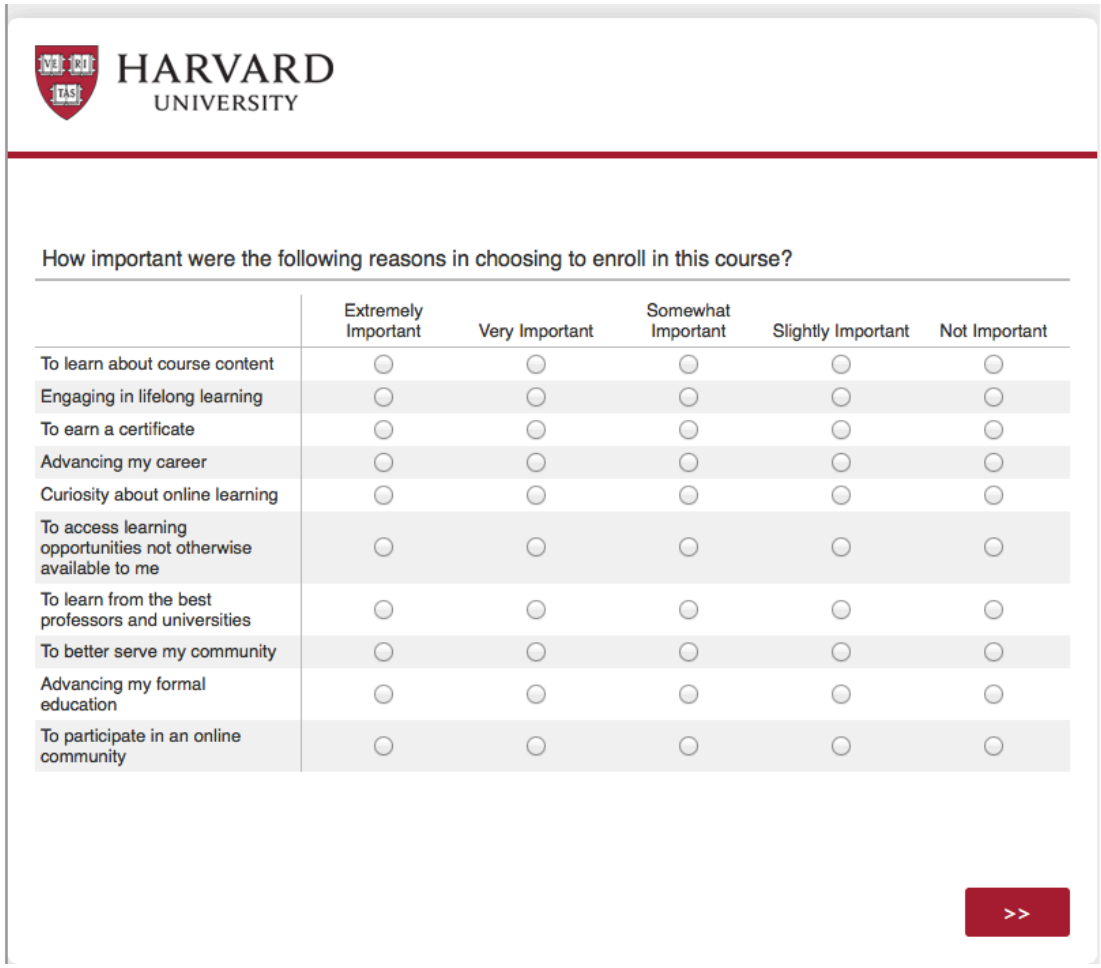


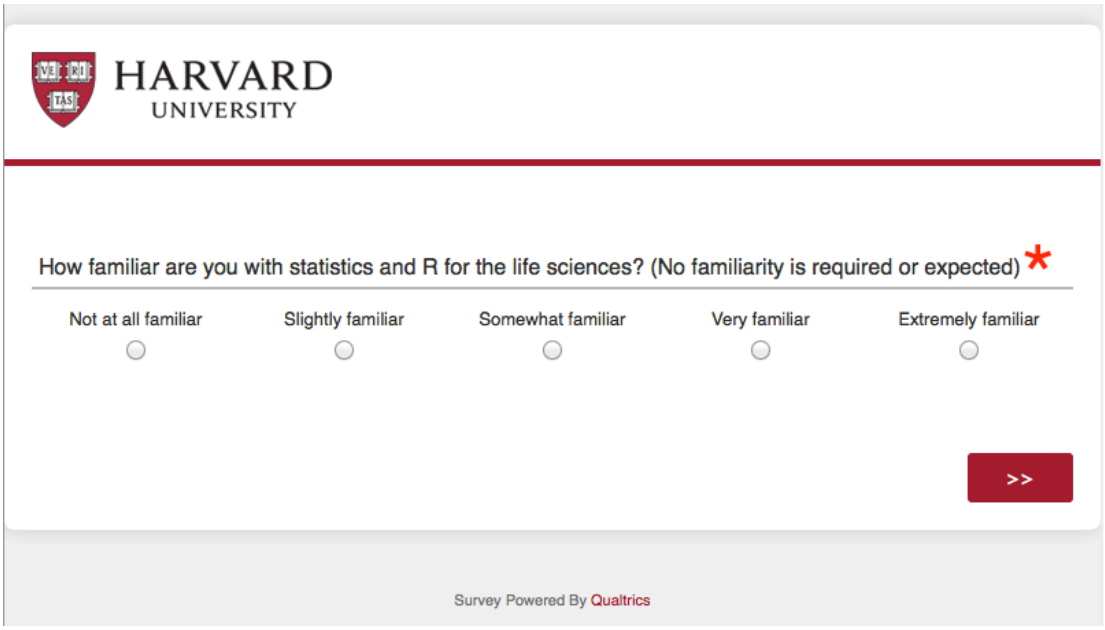
5. HarvardX Standard Pre-course Survey (example from a course called "Statistics and R for the Life Sciences")

Red asterisks, which will not be visible to survey participants, are placed next to questions that determine a value of a predicted moderator variable. These are followed by the associated coding for analysis.

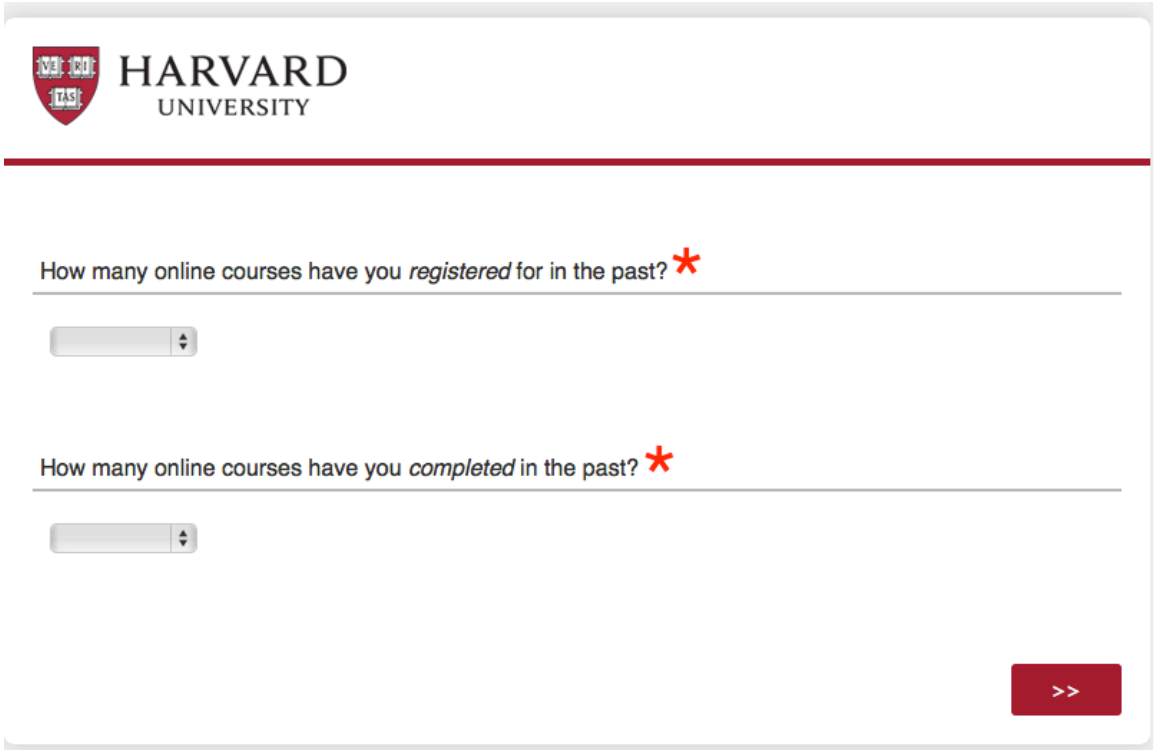


* Commitment to Complete Course = 1, 2, or 3 corresponding to the first three options in the order that they are listed





* Familiarity with Subject = 0, 1, 2, 3, or 4 corresponding to the options in the order that they are listed



* Number of Online Courses Previously Enrolled In = the number entered for the first question above

* Number of Online Courses Completed = the number entered for the second question above

6. Field Experiment-specific Survey

We distributed this survey in the last two weeks of the each of the main and supplemental experiments.

1. How important is it to you to be more active in the course than...

	Not at All Important	Somewhat Important	Important
a. 50% of your classmates	○	○	○
b. 75% of your classmates	○	○	○

2. Are you interested in seeing how your activity in the course compares to the...

	No	Somewhat	Yes
a. classmate median	○	○	○
b. classmate top quartile	○	○	○

3. How confident are you in your ability to be more active in the course than...

	Not at All Confident	Somewhat Confident	Confident
a. 50% of your classmates	○	○	○
b. 75% of your classmates	○	○	○

Answer Coding

We code a student's response to each question as a one, two, or three in order from negative to affirmative. In Table 23, we provide charts of the percentage of respondents

selecting each response and test for differences in the distribution of answers regarding the median and top quartile respectively.