

Henry Eyring, V.G. Narayanan

Performance effects of setting a high reference point for peer-performance comparison

**Article (Accepted version)
(Refereed)**

Original citation: Eyring, Henry and Narayanan, V.G. (2017) *Performance effects of setting a high reference point for peer-performance comparison*. [Journal of Accounting Research](#). ISSN 0021-8456

DOI: [10.1111/1475-679X.12199](https://doi.org/10.1111/1475-679X.12199)

© 2018 [University of Chicago](#) on behalf of the Accounting Research Center

This version available at: <http://eprints.lse.ac.uk/86732/>

Available in LSE Research Online: February 2018

LSE has developed LSE Research Online so that users may access research output of the School. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LSE Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain. You may freely distribute the URL (<http://eprints.lse.ac.uk>) of the LSE Research Online website.

This document is the author's final accepted version of the journal article. There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

Performance Effects of Setting a High Reference Point for Peer-Performance Comparison

Henry Eyring^a

V.G. Narayanan^b

January 2018

Abstract

We conduct a field experiment, based on a registered report accepted by the *Journal of Accounting Research*, to test performance effects of setting a high reference point for peer-performance comparison. Relative to providing the median as a reference point for online students to compare themselves to, providing the top quartile: damps performance for those below the median; boosts performance for those between the median and top quartile; and, in the case of outcome but not process comparison, boosts performance for those above the top quartile. We do not find that either reference point yields a greater average performance effect. However, providing the more effective reference point in each partition of initial performance yields a 40% greater performance effect than providing either reference point uniformly. Students access the online courses intermittently over the span of a year. Our effects derive from small portions of our treatment groups—5% in the case of process comparison and 26% in the case of outcome comparison—who accessed treatment and who were, on average, more active leading up to and during our intervention.

JEL Classification: C93, D91, I21, M41

Keywords: Relative performance information; reference points; performance; social comparison.

Accepted by Douglas Skinner. This paper is the final Registered Report resulting from the Registration-based Editorial Process (REP) implemented by JAR for its 2017 conference; details of the process are available here https://research.chicagobooth.edu/-/media/research/arc/docs/journal/rep_policies_jar.pdf?la=en&hash=B64B8F1D368D8300BC83898C7E5CFD0557E71312.

The accepted proposal and an Online Appendix are available here <https://research.chicagobooth.edu/arc/journal-of-accounting-research/online-supplements>.

We thank two anonymous reviewers for their thorough feedback. Dennis Campbell, Tatiana Sandino, Ian Gow, Rajesh Vijayaraghavan, and participants at two Harvard Business School seminars provided helpful comments. We also thank Paul Smith for his assistance with software development and the administrators at HarvardX, including Heather Sternshein, Rafael Irizarry, and Glenn Lopez, who generously hosted the field experiment.

^a London School of Economics and Political Science, London WC2A 2AE

^b Harvard Business School, Harvard University, Boston MA 02163

1. INTRODUCTION

We test performance effects of setting a high reference point in relative performance information (RPI). RPI shows how one is performing relative to peers and motivates performance in a variety of settings, including when performance is not tied to pay (Allcott [2011], Hannan, Krishnan, and Newman [2008], Taftkov [2013]). Theories of social comparison, reference points, expectancy, and goals could help explain performance effects of the height of reference points that are commonly displayed in RPI.¹ Insight into performance effects of RPI reference point height could guide the many government, nonprofit, and corporate administrators who use RPI (e.g., Blanes i Vidal and Nossol [2011], Hallsworth et al. [2017], Song et al. [2017]). However, empirical evidence on these performance effects is lacking.

We provide such evidence using a field experiment, based on a registered report accepted by the *Journal of Accounting Research*, in online education. Our research builds on recent field studies that suggest that the peer median reference point does not motivate above-median performers to improve (Allcott [2011], Chen et al. [2010], Schultz et al. [2007]). We extend these studies by testing whether a higher reference point, the top quartile, has a different performance effect. By comparing the median and top quartile as reference points, we assess reference points that RPI designers often provide and that RPI literature has focused attention on (Bizjak, Lemmon, and Nguyen [2011], Chen et al. [2010], Gong, Li, Shin [2011], Grote [2005]).

In our field experiment, we offer RPI with either of these two reference points, each correctly labeled as the median or top quartile, for a period of two months to students in online courses. We use data on a range of actions in the courses and a record of each time a given student accesses RPI. The field setting and our intervention do not involve explicit incentives for

¹ We study reference points that are percentiles of the peer-performance distribution. We use the term “RPI reference point height” to refer to how high a percentile of the distribution is provided as a reference point in RPI.

performance, which helps us to identify the distinct information effects of RPI reference point height.

RPI and reference points work through multiple forces to influence performance. RPI allows peer comparison and so activates innate incentives to attain a positive self-image by outperforming peers (Smith [2000], Brown et al. [2007]). A reference point instils loss aversion in those who fall short and so motivates them to work to reach the reference point (Abeler et al. [2011], Allen et al. [2017]). Expectancy theory adds a qualification on the incentives that RPI and reference points produce; a level of performance must seem attainable to motivate improvement (Atkinson [1957]). We draw on these and related theories to predict performance effects of providing a higher RPI reference point than the median.²

As planned in our registered report proposal, we assess both an average performance effect as well as performance effects in cross sections of our sample. Theory suggests that providing a higher, rather than lower, reference point will most positively affect the performance of individuals who are initially between the two alternatives. This implies a concave relationship between the positive performance effect of reference point height and an individual's initial performance.

We find the predicted concave relationship. Relative to providing the median in RPI as a reference point, providing the top quartile: damps performance for those below the median; boosts performance for those between the median and top quartile; and, in the case of RPI on an outcome measure but not a process measure, boosts performance for those above the top quartile.

To understand the mixed results for top-quartile performers, we draw on survey evidence on the outcome and process measures in our study. The outcome measure, Grade, is the

² Festinger [1954] and Smith [2000] address foundational theory regarding social comparison, Kahneman and Tversky [1979] regarding reference points and loss aversion, and Atkinson [1957] and Vroom [1964] regarding expectancy-based motivation.

percentage of problems a student has answered correctly. The process measure, Activity Level, is a weighted sum of a student's actions in the course, such as logins and video views. We find persistent (diminished) interest in outperforming peers toward higher levels of Grade (Activity Level). These responses align with social comparison theory. Outcome comparisons that reflect ability tend to encourage competition to rise well above the median (Smith [2000], Tafkov [2013]). Behavior and process comparisons tend to discourage substantial deviation from the median (Allcott [2011], Chen et al. [2010], Dolan et al. [2012]).

Our display of the top-quartile reference point for Grade could motivate top-quartile performers to further exceed it because they desire to distinguish themselves significantly by that measure. By contrast, our display of the top-quartile reference point for Activity Level might not yield such a performance response because top-quartile performers feel it is not important, and perhaps suboptimal, to be so far from the norm of a process measure. This theoretical discussion goes beyond our initial proposal to shed light on our unplanned comparison of Grade RPI and Activity Level RPI.

We did not plan tests of Grade RPI in our approved proposal, though the referee suggested them prior to the proposal's approval, because our field site did not expect to be able to provide the necessary data. We later received the necessary data. We present our planned hypothesis tests of Activity Level RPI in section 4.3 and present the same tests applied to Grade RPI in section 4.4. In section 4.5, we compare our results for RPI on each measure and discuss the comparisons.

In a planned supplementary analysis, we find that a small portion of our treatment groups accessed RPI—5.6% for Activity Level RPI and 26.09% for Grade RPI. We conduct unplanned analysis into possible reasons for low treatment access rates. Students use the courses

intermittently over the span of a year; those not active around the time of our two-month intervention would be less exposed to and might be less interested in the treatment. We find that, relative to students who did not access their course in the month leading up to our intervention, those who did were more than twice as likely to access RPI. Moreover, 21.48% (75.24%) of treatment-group students who were active on more than 10 days during our intervention accessed Activity Level (Grade) RPI.

Along with intermittent course access, another explanation for low RPI access rates seems based in the topic of RPI. Students are less likely to access Activity Level RPI than Grade RPI. Social comparison theory suggests that interest in Grade RPI might be greater because it incorporates accuracy and so reveals more about relative ability (Smith [2000], Tafkov [2013]).

The low treatment access rates in our study raise a few important considerations in interpreting our results. First, as in RPI research we draw on, our results derive from a small portion of individuals who accept the offer to view RPI and who are more engaged in the task leading up to and during performance reporting (Chen et al. [2010]). Second, limited treatment access reduces the sizes of effect estimates. For students who accessed treatment, we estimate greater effect sizes and show these in the online appendix. Third, different treatment access rates for Grade RPI and Activity Level RPI could partly explain the mixed results among top-quartile performers. We discuss in section 5 how these results could arise if treatment-group portions that access Grade RPI are inherently more competitive than those that access Activity Level RPI.

We use planned supplementary analyses beyond the topic of treatment access to further explore causes and conditions of our main results. Our survey evidence shows less confidence in one's ability to reach the top quartile than the median. Combined with expectancy theory, this

evidence suggests that low performers may respond better to the lower reference point in part because they question their ability to reach the higher reference point.

In another planned supplementary analysis, we test whether gender moderates performance responses to RPI reference points. Research shows gender differences in the performance effect of peer comparison that occurs through interaction with a high-performing peer (Lavy, Silva, and Weinhardt [2012]). We do not find gender differences in the performance effect of peer comparison that occurs through display of a high RPI reference point. We state a confidence interval for this null result and discuss possible explanations, including a lack of statistical power, in the online appendix. One possible explanation is that gender differences in response to interaction with a high performing peer depend on the identifiability of that peer and related chances to cooperate (Cross and Madson [1997], Eagly [1978]), which anonymous RPI reference points do not offer.

Our study makes three main contributions. First, we show effects of RPI reference point height that operate through the display of anonymous performance information. This speaks to the growing body of economic and psychology research on this type of information display as a tool for influencing performance and behavior. This research spans the private and public sectors and shows that RPI affects retail service, educational attainment, energy consumption, web-site content contribution, and taxpaying.³ Research on the importance of the reference point displayed along with RPI could have policy implications for a variety of corporate and other societal settings.

Second, we extend accounting research on moderating conditions of RPI's performance effects. Accounting research has addressed conditions that include financial compensation,

³ See Blanes i Vidal and Nossol [2011] regarding retail and wholesale, Azmat and Iriberry [2010] regarding education, Allcott [2011] and Schultz et al. [2007] regarding energy consumption, Chen et al. [2010] regarding web-site content contribution, and Hallsworth et al. [2017] regarding taxpaying.

anonymity, information detail, and multitasking (Hannan, Krishnan, and Newman [2008], Hannan et al. [2013, 2017], Newman and Tafkov [2014], Tafkov [2013]). We contribute to these studies by addressing the height of reference points. Given the prevalence of reference points in applications of RPI, reference point height is a salient feature to provide evidence on (Bizjak, Lemmon, and Nguyen [2011], Song et al. [2017]).

Third, analysis of the effects of RPI reference points in isolation informs theory and empirical work in a variety of other streams of research. Research on RPI-related accounting and economic mechanisms might draw from our results in a number of ways.⁴ For example, we find weak returns to RPI reference point height relative to those that other studies have shown to goal or target height in the absence of RPI (Erez and Zidon [1984], Locke and Latham [2002]). A partial explanation for the prevalence of achievable targets could be that targets often serve as reference points in RPI and, in that role, introduce some of the negative performance responses to RPI reference point height that we find.⁵ Also, tournament literature notes 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 approach to performance management—providing a lower reference point to low performers and, in the case of outcome-based performance, a higher reference point to high performers. Furthermore, evidence on the performance returns to RPI reference point height could inform prediction of the effects of supervisor discretion in setting targets that communicate RPI (Bol et al. [2010]). Lastly, RPI

⁴ For examples of such RPI-related mechanisms, see Aranda, Arellano, and Dávila [2014], Bol et al. [2010], and Murphy [2000] regarding target setting, Securities and Exchange Commission [2015] regarding disclosure and monitoring, and Gibbons and Roberts [2012] regarding contracting.

⁵ See Merchant and Manzoni [1989] for an overview of positive returns to goals and targets achievable less than 50% of the time, and the contrasting prevalence of more achievable targets in practice. Bouwens and Kroos [2011], and Leone and Rock [2002] also find highly achievable targets in practice. Merchant and Manzoni [1989] and Ioannou, Li, and Serafeim [2016] discuss rationales for highly achievable targets despite evidence of returns to target height. Aranda, Arellano, and Dávila [2014] and Bol et al. [2010] are examples of the use of targets along with RPI.

reference points, especially the median and top quartile, are commonly used in measuring corporate performance and evaluating employees.⁶ Research in those settings might use our results to account for behavioral responses to the display of RPI reference points.

2. THEORY AND HYPOTHESIS DEVELOPMENT

Economics-based research addresses functions of RPI that explain its widespread use (Gibbons and Roberts [2012] p. 67). Principals use RPI to filter out common noise from a performance measure so that it imposes less risk on an agent in an incentive contract (Banker and Datar [1989], Holmstrom [1979], Lazear and Rosen [1981]). Agents use RPI to form expectations of pay for marginal effort based on their proximity to relative-performance cutoffs in nonlinear incentive schemes (Asch [1990], Casas-Arce and Martinez-Jerez [2009], Hannan, Krishnan, and Newman [2008]).

Recent research outlines an additional role of RPI, in which RPI influences performance even in the absence of incentive contracts. Studies show generally positive performance effects of providing RPI in fixed-pay, piece-rate-pay, and no-pay contexts (Azmat and Iriberry [2010], Murthy [2010], Tafkov [2013]).

In many applications, RPI includes reference points for peer comparison, often peer-median performance (Blanes i Vidal and Nossol [2011], Chen et al. [2010]). Like RPI, reference points can influence effort above and beyond the rate of pay for marginal performance. Recent economic studies incorporate reference points in models of utility that traditionally only weigh monetary pay-off and cost of effort. In positively weighting a reference point, utility functions

⁶ See Bebhuk 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.

account for “reference-dependent preferences” (Abeler et al. [2011], Farber [2008]). Although RPI and reference points can both motivate effort, little is known regarding the effect of RPI combined with reference points of varying heights.

A number of field studies report generally positive performance effects of RPI display but do not test the performance effects of varying the height of reference points (Allcott [2011], Allcott and Rogers [2014], Azmat and Iriberry [2010], Chen et al. [2010], Schultz et al. [2007]). These studies show performance effects that depend on initial performance relative to the reference point; individuals underperforming a reference point exhibit a more positive performance response than those outperforming it. Researchers suggest that this may be due to the difficulty of achieving beyond an already high level or to a downwardly attractive power of reference points for those performing above them (Allcott [2011], Schultz et al. [2007]). A potential implication is that a higher reference point, relative to which individuals would be situated differently, would yield different performance effects.

Theories of social comparison, reference points, goals, and expectancy explain how RPI and reference points influence performance. We addressed each theory in our initial proposal and include the same insights from each here to provide context for our predictions of the performance effect of RPI reference point height.

Social comparison theory explains how RPI drives performance. RPI adds a performance incentive in the form of utility from comparing favorably to peers (Brown et al. [2007], Garcia and Tor [2007], Tafkov [2013]). This incentive cuts across a variety of contexts. RPI drives performance with or without performance-based pay, whether peers are identifiable or anonymous, and whether one’s performance is visible to others or kept private (Hannan,

Krishnan, and Newman [2008], Klar and Giladi [1997], Tafkov [2013], Xiao and Lucking [2008]).

Reference points, or levels for comparison to, follow principles of behavioral economics in influencing performance (Heath, Larrick, and Wu [1999]). Individuals feel loss aversion to falling short of a reference point and so try to reach it. Economic models of utility that account for reference points incorporate a loss aversion term, and a lab study shows that 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]). Reference points for effort provision can also arise separately from financial contracts. For instance, runners exert effort near the end of a marathon to finish a few seconds before a round number time (Allen et al. [2017], Markle et al. [2015]). Whether a reference point is introduced, as in performance management interventions like ours, or arises naturally, as with a round number, researchers use loss aversion to explain its ability to boost performance (Abeler et al. [2011], Allen et al. [2017]). We take this same theoretical approach.

Expectancy theory states that motivation to achieve a level of performance depends on its “expectancy,” or perceived attainability (Atkinson [1957], Lawler and Suttle [1973], Vroom [1964]). Goal theory similarly notes that perceived attainability is fundamental to a goal’s ability to motivate performance (Erez and Zidon [1984], Locke, Motowidlo, and Bobko [1986]). We consider in our hypotheses how the level of expectancy of the RPI reference points we test could influence their performance effects.

An additional motivator present in our study and in a variety of corporate and public-sector settings is a visual indication of approval for performing well relative to a peer reference point (Allcott [2011], Campbell [2006], Vanek Smith [2015]). Examples from these settings

include red and green colors to signal low and high performance, respectively, and smiley faces to signal high performance. We adopt smiley faces, as used in field experiments from psychology and economics, to test the performance effects of RPI reference points when outperforming them is visually congratulated (Allcott [2011], Schultz et al. [2007]).

Hypothesis 1 through hypothesis 3c are stated here exactly as in our approved proposal. Hypothesis 1 regards the validity of our intervention as a performance management tool. We state our hypotheses in the alternative form.

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

We incorporate theories of RPI and reference points in developing Hypotheses 2a-3c, which relate to RPI reference point height. While we do so more succinctly than in the approved proposal, in line with feedback from the referee and editor, the theories lead us to the same predictions.

We predict a concave relationship between an individual's initial performance and the performance effect of providing the top quartile, rather than median, reference point. We address the predicted concavity among partitions of initial performance, from low to high.

We predict a negative effect of providing the top-quartile, rather than median, reference point to initially below-median performers. Reference points offer lower marginal utility to effort when farther away, and so the top-quartile reference point would offer this group lower marginal utility than would the median (Abeler et al. [2011], Heath, Larrick, and Wu [1999], Kahneman and Tversky [1979], Markle et al. [2015]). The higher reference point would also hold lower

expectancy, a key element of motivation (Atkinson [1957]). From a social comparison theory standpoint, upward comparison to a distant level of high performance has been shown to discourage effort (Rogers and Feller [2015]). Research on tournaments and rank-based pay similarly suggests that individuals are discouraged to the point of giving up when they feel that a nonlinearly rewarded level of performance is too high (Bandiera, Barankay, and Rasul [2013]).

On the other hand, reference point research implies one advantage of the higher reference point among below-median performers. Individuals who surpass the median might be drawn higher still by the higher reference point (Abeler et al. [2011], Heath, Larrick, and Wu [1999], Reno, Cialdini, and Kallgren [1993]). The balance of theory, though, suggests that the higher reference point will be less motivating than the lower reference point among below-median performers.

For those initially between the two reference points, providing the top-quartile, rather than median, reference point is more likely to boost performance. The top-quartile reference point is more attainable, and so holds a higher degree of expectancy, for the 50th—75th-percentile performers than for below-median performers. The top-quartile reference point also rewards effort for all individuals in this partition in the form of reduced feelings of loss from underperforming the reference point. The median reference point, by contrast, offers such utility only to those who fall below the median (Abeler et al. [2011]). We note that many studies show the relevance of the median as a reference point and we can find fewer that address the top quartile (Dolan et al. [2012], Larrick, Burson, and Soll [2007]). Other than the related possibility of greater innate interest in and desires to outperform the median reference point, theory suggests a positive performance effect of providing the higher, rather than lower, reference point to initially 50th—75th-percentile performers.

In the top-quartile of initial performance, a few forces could cause the higher reference point to be more motivating than the lower. Top-quartile performers are more likely to fall beneath the higher reference point. The prospect and instance of that event might yield stronger performance incentives. Also, to the extent that RPI reference points carry downward attractive power, a higher reference point might act as a bulwark to mitigate a performance decline (Schultz et al. [2007]). However, evidence of a downward attractive force of RPI reference points is mixed (Allcott [2011], Chen et al. [2010]).

A few forces could, by contrast, cause the higher reference point to be less motivating than the lower in the top-quartile of initial performance. The higher reference point reveals to top-quartile performers that they are in the right tail of the performance distribution. This could cause them to feel complacent or to worry that their behavior is suboptimal (Schultz et al. [2007]). Also, if the median generates more interest in social comparison than the top quartile, this could lead to greater effort to perform well by the comparison (Dolan et al. [2012], Larrick, Burson, and Soll [2007]). We predict a positive effect of providing the top-quartile, as opposed to median, reference point among initially top-quartile performers despite some theory to the contrary. The more uniform predictions from related theories are for a performance benefit of reference point height among individuals initially performing between two alternatives. We predict that these individuals will respond more positively on average to reference point height than will those initially above the higher reference point.

To address how our study differs from studies of goal difficulty we add theory beyond that included in our approved proposal. Research on goals—explicitly stated objects or aims of an activity—generally reports that the more difficult the goal, the higher the performance, even to the point at which less than 25% of individuals achieve the goal (Erez and Zidon [1984],

Locke and Latham [2002], Locke, Motowidlo, and Bobko [1986]). We test for an average performance effect of RPI reference point height at similar levels of attainability, but do not predict its sign because RPI reference points reveal information beyond the information in goals.

In particular, RPI reference points reveal information on relative performance and social norms. As we noted earlier in this section, a comparison to high-performing peers might cause low performers to feel discouraged and high performers to feel complacent or that they are deviating in a suboptimal way from social norms. These possibilities drive our predictions of a concave relationship between initial performance and RPI reference point height, whereas studies show positive returns to goal difficulty across spectrums of initial performance and ability (Erez and Zidon [1984], Latham, Seijts, and Crim [2008]). A mix of negative and positive effects of RPI reference point height might yield no or even a negative average effect, unlike the positive average effect found for goal difficulty.

Hypotheses 2a-c predict the effect of providing the top-quartile, rather than median, reference point in each partition of initial performance addressed above. We predict opposing effects among partitions that may balance out or yield a net effect. As a result, we test for but do not predict the net effect.⁷

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.

⁷ The approved registered report proposal similarly noted that we would test for but not predict the occurrence or sign of a net effect of providing the top-quartile, rather than median, reference point. We included in that proposal a hypothesis 4 right after that statement with the alternative of an effect and a null of no effect. We agree with subsequent feedback that stating a hypothesis without a prediction of an effect's occurrence or sign was a deviation from the regular meaning of a hypothesis. We thus now only state that we will test for a net effect despite not making a prediction of the net effect.

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 the concave relationship between initial performance and returns to a higher reference point that H2a-c imply. H3a predicts a more positive performance effect of providing the higher, rather than lower, reference point for those initially between than outside the alternative reference points. H3b (H3c) predicts a more positive effect in the in-between partition than in the lower (higher), which helps 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.

3. FIELD SETTING AND DATA

3.1 Field Setting

In 2012, The Massachusetts Institute of Technology and Harvard University jointly founded edX, a nonprofit organization offering free online courses, assessments, and certificates for higher education. 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. All instruction occurs online. Course topics range from literature to statistics, and courses 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.

3.2 Experiment Design

Our study's experiment design consists of a control group, which received no RPI display, and two treatment groups. One treatment group received an RPI display with the peer median reference point and the other received an RPI display with the peer top-quartile reference point.

A student's RPI display showed his or her performance relative to a single reference point, correctly labeled as either the median or top quartile.

We provided students in each of the two treatment groups with access to RPI for a period of two months. We delivered the RPI using weekly emails that led a student to his or her RPI display. To increase exposure to the intervention, we placed a link reading "Check your progress" within the course platform. The link directed control-group students who clicked it to the default HarvardX progress chart for the course, which showed completion status of individual assignments and no RPI. The link directed treatment-group students who clicked it to their RPI, below which sat a link to the default HarvardX progress chart for the course. We updated the displays daily. Appendix B contains example displays.

We initially proposed using Activity Level, as opposed to Grade, as the measure of performance. Our choice of the former was driven by data availability; HarvardX could not provide the necessary data at the time our proposal was approved.

The referee, though, encouraged us to test RPI regarding Grade. The referee noted that individuals might desire to rise to the high tail of the distribution of an outcome measure, like Grade, that reflects ability but not of a process measure, like Activity Level, that focuses more on effort. We describe related theory in section 4.5. After the proposal's approval, we gained access to data that allowed providing RPI on the outcome measure Grade. We ran the experiment as approved, with Activity Level RPI, and then modified, with Grade RPI. We refer to the former as the "main experiment" and the latter as the "supplemental experiment."

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 in the

course, 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.

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

Grade represents the percentage of the course's total problems that the student answered correctly. Students can complete problems in any order. This facilitates flexible, modular learning. The low mean Grade for students who have attempted problems does not reflect equally low accuracy, but rather students selecting which problems they will attempt and their imperfect accuracy on those attempts. We used Grade rounded to the nearest whole percent (e.g., 82% rather than 82.12%), and Activity Level rounded to the nearest integer (e.g., 19 rather than 18.74), for display during the experiments and in the analysis.⁸

Appendix A contains a full list of variable definitions. The dependent variable in the main experiment is Δ Activity Level and, in the supplemental experiment, Δ Grade.⁹ We use the other variables in the same manner in the supplemental experiment as in the main experiment.

⁸ We did not specify rounding in the accepted proposal, but the rounding occurred before the experiments were run and prior to analysis. We gained access to Grade after the proposal's acceptance and only in a rounded format—the format that our host courses use in reporting Grade to students. We rounded Activity Level in order to use graphical display software that only accepted integers. Performance measures are often shown as an integer or integer percentage, and so our study aligns with practice in that respect (e.g., Bizjak, Lemmon, and Nguyen [2011], Chen et al. [2010], Grote [2005]).

⁹ We calculate Activity Level and Grade at the end of both experiments 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

4. ANALYSIS

4.1 Analytical Approach

Our study draws both from null hypothesis significance testing (NHST) and Bayesian analysis. We conduct NHST for each hypothesis. We calculate Bayes factors when we fail to reject a null hypothesis. Bayesian analysis makes use of effects that are not significant at traditional levels by explaining how much more probable we can expect a significant relation to be than we could before the realization of the data.

We apply guidance from research on RPI and online settings in selecting our sample. One empirical challenge is zero-inflation from the large percentage of students who enroll and then do not participate in these courses (Lamb et al. [2010]). We exclude these students from our study. Of those who access the course, a large number do not try graded content. We exclude these students in the supplemental experiment to avoid zero-inflating the displayed standard of performance. Chen et al. [2010] make a similar sample restriction to the active portion of an online community.

The influence of a small number of outliers who use an online course approximately 10 times more than the 99th percentile poses another empirical challenge (Lamb et al. [2010]). We winsorize values for Activity Level at the 99th percentile, as specified in our proposal, to ensure that a small number of extreme outliers do not drive or offset results.¹⁰ For consistency, we perform the same winsorization on any components of Activity Level used as dependent

this paper to show how Δ Activity Level and Δ Grade correlate. However, by including Activity Level and Grade in the descriptive statistics for each experiment, we show how these performance measures generally correlate in the host courses.

¹⁰ Our results hold at the same level of significance when we exclude individuals above the 99th percentile of Activity Level, or its components when used as dependent variables, rather than winsorizing at the 99th percentile.

variables in unplanned analyses. Winsorization is not necessary for Grade, which is capped at 100.

We cluster standard errors at the student level to correct for autocorrelation from students who enroll in more than one host course for the experiment. Any students present in more than one course are included in the same experimental group in all courses. Their experimental group membership is, as with all students, set through random assignment.

4.2 Descriptive Statistics

Tables 1 and 2 show the sample selection and descriptive statistics for the main and supplemental experiments, respectively. The courses attract individuals who are on average in their thirties. 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.

Turning to performance, the average Activity Level at the beginning of the main experiment was 96.5. An example student with that activity profile who had a typical distribution of activity among course elements would, at the experiment's start, have been active in the course on three days, watched 15 videos, made one discussion forum post, taken six other actions in the discussion forum, tried nine problems, and made 160 clicks on course content. Activity Level rose to an average of 117.73 at the experiment's end, a roughly 22% increase in the weighted component measures.

Grade stood at an average of 19.02 at the supplemental experiment's start and rose by 2.63 to an average of 21.65 at the experiment's end—a roughly 14% increase. A student's Grade starts at zero and rises with each question correctly answered. The maximum possible Grade is 100.

Tables 3 and 4 show a planned supplementary analysis in which we find limited treatment access. 5.6% (26.09%) of the treatment groups in our experiment regarding Activity Level (Grade) RPI access the RPI. We do not find that the number of students who return to access RPI more than once depends on the reference point that we provide. The courses are self-paced, and students access the courses intermittently; the average number of days between course access is 16, or over one-fourth of the length of our intervention. We note in section 1 that unplanned analyses show that those who are more active in the courses leading up to and during the intervention are more likely to access treatment. Treatment access is also more common for Grade RPI than Activity Level RPI. We describe implications of treatment access in section 1 and section 5.

Our hypotheses compare performance among cells of initial performance and assigned treatment. Table 5 (table 6) displays these cells in a grid for Δ Activity Level (Δ Grade). These tables show the values of the dependent variables in cells that we compare in our hypothesis tests.

Figures 1 and 2 depict the distributions of our dependent variables by initial performance and reference point. Below the median of initial performance, individuals shown the median reference point perform best. Between the median and top quartile of initial performance, individuals shown the top-quartile reference point perform best. For initially top-quartile performers in the case of Activity Level RPI, those shown the median reference point perform best. For initially top-quartile performers in the case of Grade RPI, those shown the top-quartile reference point perform best. These graphs help to visualize the dependent variable descriptive statistics prior to formal hypothesis tests.

We address our planned hypothesis tests in section 4.3. In section 4.4, we apply those same hypotheses to the unplanned, supplemental experiment by using Grade as the measure of performance instead of Activity Level. In section 4.5, we report additional analyses based on planned supplementary tests as well as unplanned tests.

4.3 Tests of Planned Hypotheses

Table 7 shows tests of our planned hypotheses. We conduct the tests using a planned OLS specification and also using an unplanned cell-means specification that yields the same results, i.e. the same rejections of null hypotheses with the same estimate magnitudes significant at the same levels. We include the cell-means specification in the main text in response to feedback that we should more clearly link cells of treatment and initial performance to hypotheses and results. We show the OLS specification in the online appendix.

The first row of table 7 shows evidence to support H1, that providing RPI boosts performance. With H2a, H2b, and H2c, we predicted that providing the top-quartile, rather than median, reference point would damp performance for those initially below the median, boost performance for those initially between the reference points, and boost performance for those initially above the top quartile, respectively. We show evidence to support H2a (H2b) in the second (third) row of table 7. The evidence in the fourth row does not support H2c. We calculate a Bayes factor of over 20 that provides a strong indication, by the standards of Kass and Raftery [1995], that providing the top-quartile, rather than median, reference point damps performance for those initially above the top quartile.¹¹ In section 4.5, we discuss how our survey evidence and theory related to interest in outperforming peers could help to explain this result.

¹¹ We calculate Bayes factors for the main hypothesis tests when the tests yield a null result. These are referenced in the main text and shown in the online appendix.

The fifth row of table 7 supports the prediction from H3a of a concave relationship between initial performance and the positive performance effect of displaying the higher, rather than lower, reference point. Specifically, the performance effect is more positive for those initially between the reference points than outside them. The sixth row of table 7 supports H3b, that the performance effect is more positive for those initially between the reference points than below the median. The seventh row of table 7 supports H3c, that the performance effect is more positive for those initially between the reference points than above the top quartile. Figure 3 illustrates this concave relationship. The tendency of individuals between the median and top quartile to respond competitively to seeing the higher, rather than lower, reference point while those on either side tend not to produce a hill shape.

We do not find an average performance effect of providing the higher, rather than lower, reference point. This test is shown in the last row of table 7. The associated Bayes factor of 3.79 warrants updating a prior of no relationship in favor of the model's estimated negative effect to a level of 3.79 times the probability before the realization of the data.

4.4 Tests of Planned Hypotheses Applied to the Unplanned, Supplemental Experiment

Table 8 shows tests of our planned hypotheses that we apply to the unplanned, supplemental experiment by using Grade, rather than Activity Level, as the measure of performance. The first row of table 8 shows that providing RPI boosts performance. The second, third, and fourth rows show results directionally in line with our predictions for H2a, H2b, and H2c, that providing the top-quartile, rather than median, reference point would damp performance for initially below median performers, boost performance for those initially between the reference points, and boost performance for those initially above the top quartile, respectively. The results of the test for H2b

are not statistically significant. The Bayes factor of 2.04 warrants updating the odds of the hypothesis as stated to be just over twice as likely as before the realization of the data.

The fifth row of table 8 supports the prediction from H3a of a concave relationship between initial performance and the positive performance effect of displaying the higher, rather than lower, reference point. The sixth row of table 8 supports H3b, that the performance effect is more positive for those initially between the reference points than below the median. The seventh row of table 8 shows a result directionally in line with H3c, that the performance effect is more positive for those initially between the reference points than above the top quartile. The result is not statistically significant. The associated Bayes factor of 1.93 for H3c indicates that the predicted relationship is just under twice as likely relative to the null of no relationship in light of the data. Figure 4 illustrates this concave relationship. The tendency of individuals above the median to respond competitively to seeing the higher, rather than lower, reference point while those below tend not to produces a plateau shape.

We do not find an average performance effect of providing the higher, rather than lower, reference point. This test is shown in the last row of table 8. The associated Bayes factor of 2.79 warrants updating a prior of no relationship in favor of the model's estimated negative effect to a level of 2.79 times the probability before the realization of the data.

4.5 Additional Analysis

To understand drivers and conditions of the main effects, we conduct additional analyses. We note before each analysis whether it is a planned supplementary analysis or an unplanned analysis. In the main text, we provide an overview of each analysis. We refer the reader to the online appendix for further detail.

We use a planned supplementary analysis to address differences in performance effects by gender. Studies of social interaction point to gender as a determinant of cooperativeness (Cross and Madson [1997], Eagly [1978]). Recent evidence suggests that females are more prone to cooperate and that this leads them to benefit more than males from interaction with high-performing peers (Lavy, Silva, and Weinhardt [2012]). We do not find evidence that females similarly benefit more from comparison to a high reference point of peer performance. This null result is consistent with cooperation as a mechanism for gender differences in response to peer comparison. An anonymous reference point does not identify a particular high-performing peer and for that reason might not lead to differential cooperation and resulting performance effects for females and males. In the online appendix, we state a confidence interval for this null result and discuss other possible explanations including a lack of statistical power.

We also conduct a planned supplementary analysis of interest in outperforming peers. Table 9 shows that respondents deem it less important to be above the top quartile than the median by activity in a course. This aligns with research that shows a tendency to conform to normal behavior and processes (Dolan et al. [2012], Schultz et al. [2007]). The results in table 10 do not show that respondents deem it less important to achieve a grade above the top quartile than above the median. This aligns with research that shows a common desire to be of very high intellectual ability relative to peers (Smith [2000], Tafkov [2013]).

This analysis offers a possible explanation for our mixed results among top-quartile performers. Individuals who see that they have passed the top quartile of Grade might climb higher because of their persistent interest in outperforming peers past high levels of Grade. Individuals who see that they have passed the top quartile of Activity Level, on the other hand,

might not respond with greater effort because it is not as important to them to be in such a high portion of the distribution of that measure.

In a third planned supplementary analysis, we test confidence in the attainability of the two reference points. Tables 9 and 10 show significantly greater confidence in ability to reach the median than the top quartile in the case of both activity and grades in a course. Part of the reason that those below the median improve more if shown that lower reference point could be that it holds greater expectancy from their standpoint.

We use an unplanned analysis to assess the source of the improvement in Grade in the supplemental experiment. We show in the online appendix that Grade RPI led to a higher quantity of problems attempted but not a statistically significant increase in the accuracy of problem attempts. This suggests an effort, as opposed to aptitude, mechanism for the RPI's performance effect. Theories of reference-dependent preferences for effort provision align with this observed effort-based mechanism for the performance effects of Grade RPI (Abeler et al. [2011]).

5. DISCUSSION

Our study contributes most directly to a growing body of literature on the distinct effects of RPI, or those separate from pay for or visibility of relative performance (Blanes i Vidal and Nossol [2011], Chen et al. [2010], Taikov [2013]). While applications of RPI often include a reference point for peer comparison, little to no empirical evidence has established the effects of RPI reference point height (Allcott [2011], Chen et al. [2010]).

We explain in the abstract and in the introduction and analysis sections that small portions of our sample accessed the treatment. As in reference point research that we draw on,

our results derive from those who are more engaged in a task leading up to and during performance reporting (Chen et al. [2010]). We estimate larger effect sizes for those who access RPI and we show these in the online appendix.

Treatment access could also partly explain the different results for top-quartile performers when we show RPI on different measures. Theory suggests that social comparison will be more intense the more that RPI reveals about relative ability (Smith [2000]). Competitive types might be more prone to view Grade RPI than Activity Level RPI to the extent that the former reveals more about relative ability. The portions of our treatment groups that view and respond to RPI could be more competitive by nature in the case of Grade RPI. This could occur at the same time as interest in outperforming peers generally persists past higher percentiles of peer Grade than peer Activity Level, consistent with our survey results. Either of these occurrences or their combination could help explain why top-quartile performers respond competitively to the top-quartile reference point in the case of Grade but not of Activity Level RPI.

Viewed collectively, our results among partitions of initial performance offer some of the first evidence that the performance effect of providing a relatively high RPI reference point depends on initial performance. We map out the concave shape of performance returns to RPI reference point height along the axis of initial performance. Neither of the reference points we test, the median nor top quartile, has a more positive performance effect on average. Our results suggest that the optimal approach to providing RPI reference points is 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 as dependent on his or her initial performance and the measure of performance, the effect on Δ Activity Level (Δ Grade) is an increase of 5.09 (1.52)

from a baseline Activity Level (Grade) of 96.5 (19.02). By contrast, when we show students the median reference point in a one-size-fits-all approach, the effect on Δ Activity Level (Δ Grade) is an increase of 3.63 (1.06). For Activity Level (Grade) RPI, the customized approach yields a roughly 40% (43%) larger effect.

Our study shows a different response to RPI reference point height than responses other studies have shown to goal difficulty or social interaction with high performers. Studies generally show net positive performance effects of a high standard for comparison in the form of a goal or a high performing peer to interact with.¹² We do not find a net positive effect of a similarly high RPI reference point. This null result may be due to a lack of statistical power. Alternatively, RPI reference points may operate differently than goals and social interaction and so yield a different effect.

Specifically, beyond information inherent in a goal, a high RPI reference point reveals information on relative performance and social norms. As we describe in section 2, this information could discourage low performers and cause high performers to feel that they should regress toward social norms (Rogers and Feller [2015], Schultz et al. [2007]). Also, RPI reference points do not directly manipulate one's peer group and so may not affect learning, networking, and cooperation, which studies of social interaction with a high performing peer use to explain positive performance effects (Hanushek et al. [2003], Lavy, Silva, and Weinhardt [2012], Lyle and Smith [2014]).

While we study RPI reference points in isolation, our results may inform research in a variety of RPI-related areas. For example, targets often either explicitly contain RPI or allow inferring one's relative performance and comparison to a reference point (Aranda, Arellano, and

¹² See Locke and Latham [2002] for a review of goal literature's findings on goal difficulty, and see Hanushek et al. [2003], Hoxby and Weingarth [2006], Lavy, Silva, and Weinhardt [2012], Lazear [2001], and Lin [2010] for evidence of effects of interaction with high-performing peers.

Dávila [2014], Bol et al. [2010], Merchant and Manzoni [1989], Murphy [2000]). We find no performance benefit on average from setting a high RPI reference point. This offers one possible explanation for the prevalence of targets in practice that are achieved by the majority of the population most of the time (Merchant and Manzoni [1989]). Through its role as an RPI reference point, we would not expect a higher target to yield better performance.

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 [2005], Bizjak, Lemmon, and Nguyen [2011]). While the disproportionate prevalence of earning and executive pay targets above the peer median has drawn widespread criticism (Bizjak, Lemmon, and Nguyen [2011]), we find behavioral responses that suggest relative performance targets set above the peer top quartile might motivate performance for individuals in the top half of the distribution. Future research could weigh this dynamic along with financial and career concerns in assessing the value of high targets for performance and compensation.

The variation in effects by initial performance also offers insight for research on supervisor discretion, tournaments, and employee evaluation. Supervisors are subject to biases and political pressures in setting targets and evaluating performance (Bol et al. [2010], Bol [2011]). If this lowers or heightens standards for peer performance comparison, our results can help to identify the group of initial performers whose performance is most likely to benefit from the resulting height of the reference point. Tournament literature shows difficulty in using tournament incentives to motivate very low and high performers, and our results suggest that providing a customized RPI reference point might motivate these groups. Finally, 29% of corporations in a Corporate Executive Board survey reported using forced-curve employee rankings for performance management (McGregor [2013]). This type of ranking system

sometimes involves the median and top quartile as reference points for peer-performance comparison (Grote [2005]). Our study raises behavioral considerations for selecting such reference points. If managers have to choose one reference point or the other, they could pick the reference point that motivates the group they feel is most important. Our results suggest that it is more effective to customize the reference point based on an individual's initial performance.

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 alternatives. Furthermore, we address the moderating role of performance-measure type by testing RPI on a process measure and on an outcome measure.

We find that the effect of providing a relatively high reference point in RPI depends on one's initial performance. We test the peer top quartile and the peer median as alternative reference points. The effect of providing the higher, rather than lower, reference point is concave in initial performance. The effect is negative among below-median performers and positive among individuals initially between the alternative reference points. In the case of an outcome-based performance measure, the effect 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 customizing the reference point based on an individual's initial performance is preferable.

Managers and regulators can incorporate these results when selecting RPI reference points to yield desired behavior. This is pertinent given the growing role that RPI reference points play in such settings as retail, education, energy consumption, and taxpaying. The results

also describe incentives resulting from peer comparison and so help in understanding effects of performance evaluation and reporting.

Appendix A: Variable Definitions

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, rounded to the nearest integer: 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	The percentage of the total problems in the course that a student has answered correctly, rounded to the nearest integer percentage
Δ Grade	Grade at the experiment's end minus Grade at the experiment's beginning
Problem Attempts	The number of times that a student entered an answer to any problem
Δ Problem Attempts	Problem Attempts at the experiment's end minus Problem Attempts at the experiment's beginning
Problem-Attempt Accuracy	The percentage of problems that a student attempted during the experiment, if any, answered correctly
Dependent Variable Components	Description
Video Views	The number of times that a student started watching a video
Problem Attempts	As defined in the dependent variables section
Forum Posts	The number of posts that a student made in discussion forums
Other Forum Actions	The number of actions (e.g., voting for a post, responding with a comment to a 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	All actions in the course that are recorded electronically; these include logins, video views, problem attempts, forum posts, other forum actions, and clicks on course material
Independent Variables	Description
Control	An indicator variable equal to one if the individual is assigned to the control group, which does not receive RPI
Median Reference Point (RPI_M)	An indicator variable equal to one if the individual is assigned to the treatment group that receives RPI with the peer median reference point for Activity Level (Grade) in the main (supplemental) experiment

Top-quartile Reference Point (RPI_T)	An indicator variable equal to one if the individual is assigned to the treatment group that receives RPI with the peer median reference point for Activity Level (Grade) in the main (supplemental) experiment
Relative Performance Information (RPI)	An indicator variable equal to one if either RPI_M = 1 or RPI_T = 1
Moderator Variables	Description
Initially Below Median	An indicator variable equal to one if the individual's Activity Level (Grade) was less than or equal to the median reference point in the main (supplemental) experiment
Initially Third Quartile	An indicator variable equal to one if the individual's Activity Level (Grade) was greater than the median and less than the top quartile in the main (supplemental) experiment
Initially Top Quartile	An indicator variable equal to one if the individual's Activity Level (Grade) was greater than or equal to the top quartile in the main (supplemental) experiment
Gender	An indicator variable equal to one (zero) if the individual chose Female (Male) as their gender in the course registration process
Developed Country	An indicator variable equal to one if the individual reported his or her country of residence and the country is of UN Developed Nation status, and equal to zero if the individual reported his or her country of residence and the country is of UN Developing Nation Status
Level of Education	An indicator variable equal to one if the individual reported that he or she holds a bachelor's or higher degree
Age	The individual's age, if any, that he or she reported during registration, truncated at 5 and 100
Descriptive Variables	Description
Familiarity With Subject	Response to survey question, "How familiar are you with [course name]?" 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 browse 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 <i>registered</i> for in the past?”
Number of Online Courses Previously Completed	Response to survey question, “How many online courses have you <i>completed</i> in the past?”

Appendix B: Sample Experiment Instruments

Example Email with Link to RPI Display

From: HarvardX <noreply@qemailserver.com> on behalf of HarvardX <noreply@qemailserver.com>
Date: Friday, 8 April 2016 at 20:25
To: Jane Doe <username@domain.com>
Subject: Your Personalized Activity Comparison Graph for PH525.2x

Dear Learner,

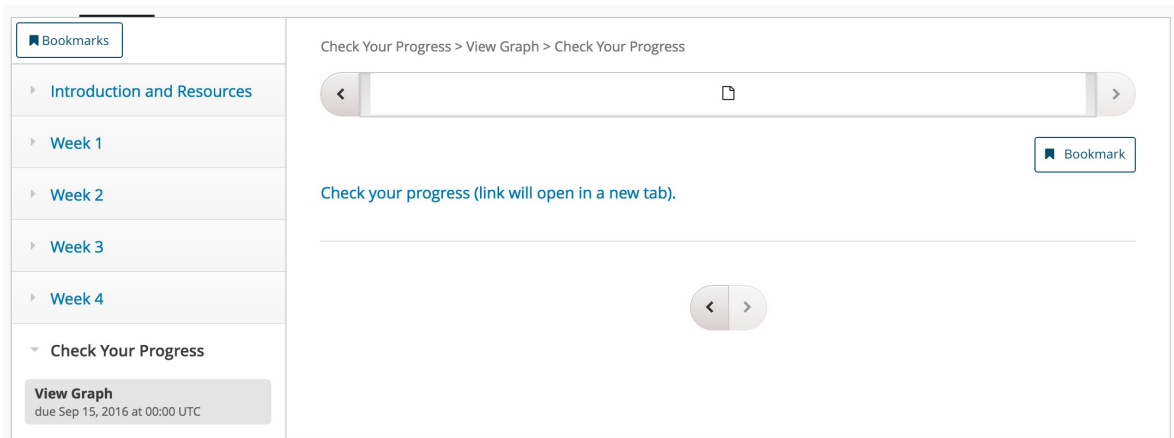
We've built a personalized comparison graph of your activity in *Data Analysis for the Life Sciences 2*. The graph will self-update daily as you do more in the class. Your most up-to-date graph will always be at this link: [see your graph](#).

Sincerely,
The Team at HarvardX

These emails will stop arriving when the course ends. If you prefer to stop receiving them immediately, follow this link: [Click here to unsubscribe](#).

In the supplemental experiment, the word “Grade” replaces any reference to “Activity.”

Example Link in Course to RPI Display



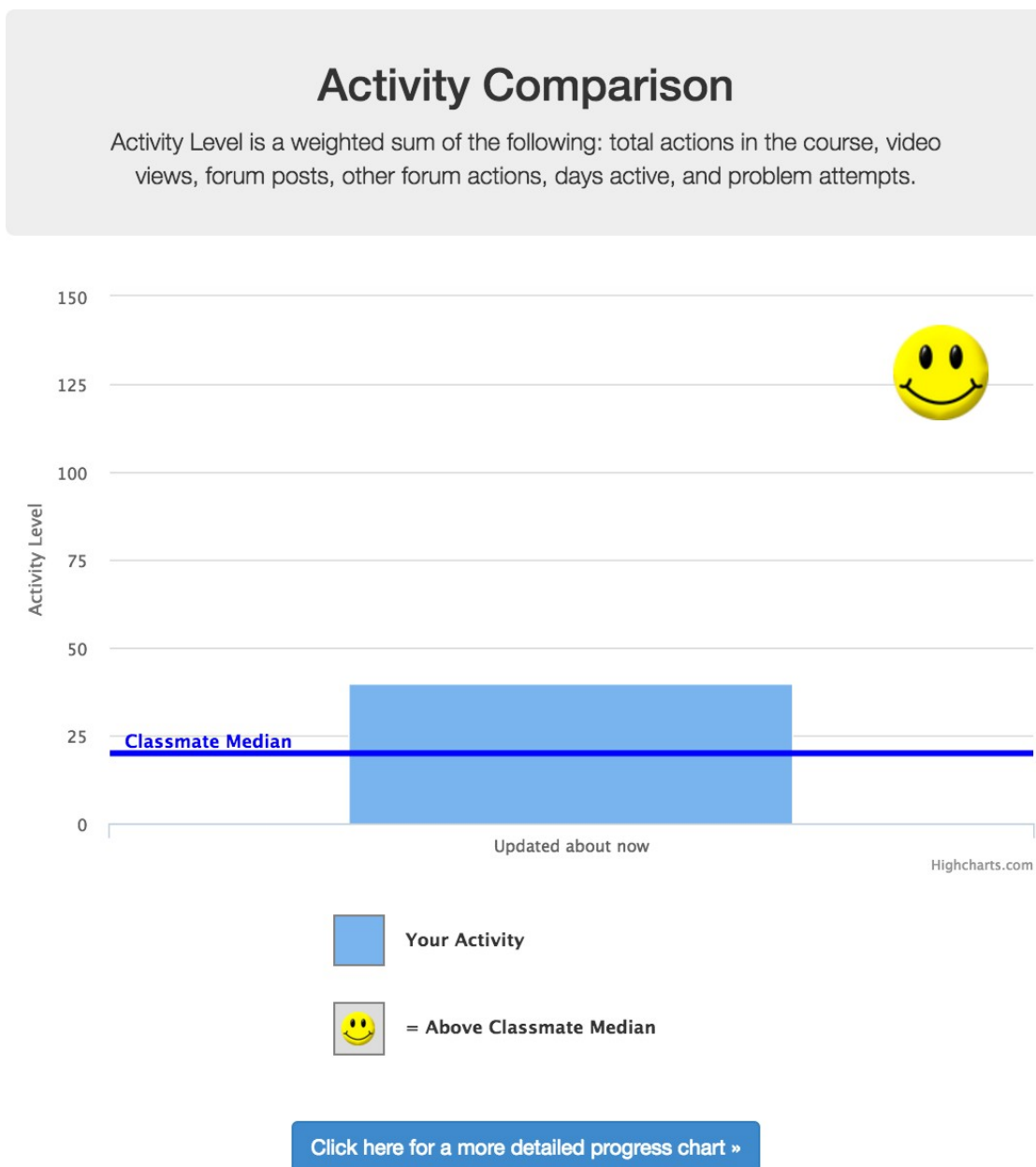
The screenshot shows a course navigation interface. On the left is a sidebar menu with a 'Bookmarks' tab at the top. Below it are links for 'Introduction and Resources', 'Week 1', 'Week 2', 'Week 3', 'Week 4', and 'Check Your Progress'. Under 'Check Your Progress', there is a 'View Graph' button with the text 'due Sep 15, 2016 at 00:00 UTC'. The main content area has a breadcrumb trail 'Check Your Progress > View Graph > Check Your Progress'. Below the breadcrumb is a horizontal progress bar with left and right navigation arrows. To the right of the progress bar is a 'Bookmark' button. Below the progress bar is a text link: 'Check your progress (link will open in a new tab)'. At the bottom of the main content area is another horizontal progress bar with left and right navigation arrows.

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.

RPI Displays

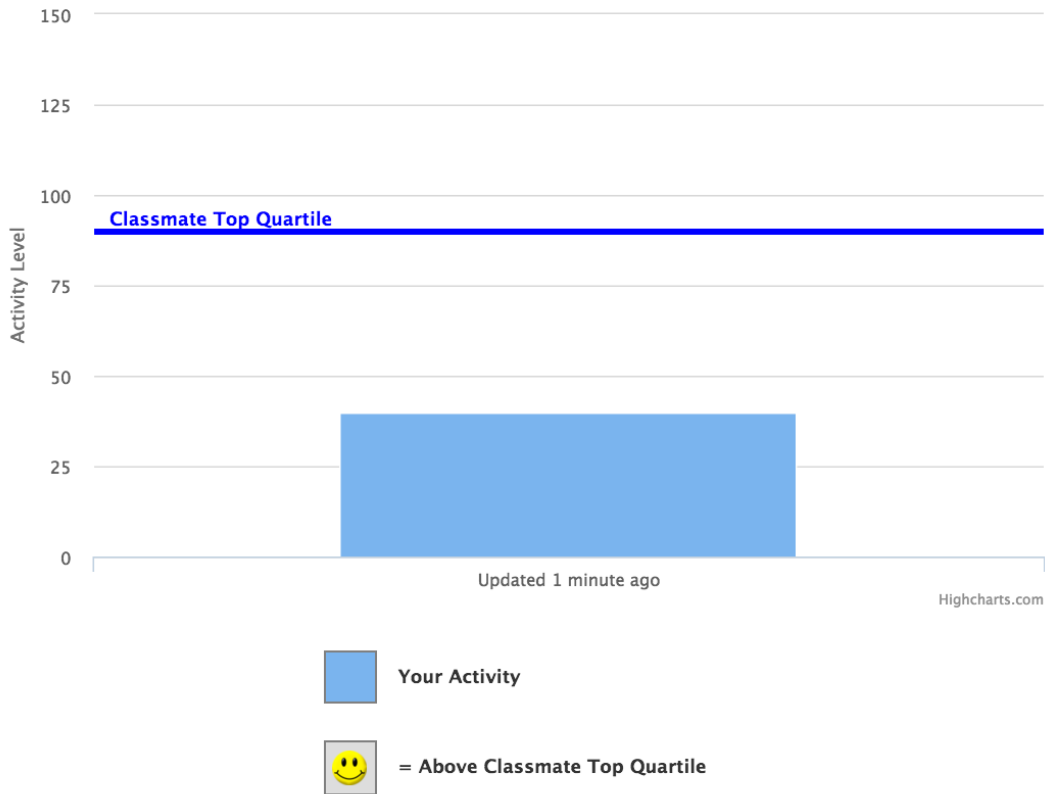
Median Reference Point



Top-Quartile Reference Point

Activity Comparison

Activity Level is a weighted sum of the following: total actions in the course, video views, forum posts, other forum actions, days active, and problem attempts.



[Click here for a more detailed progress chart »](#)

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, “Grade” replaces any reference to activity or “Activity Level,” and the annotations read “classmates who've attempted a problem” rather than “classmates.” The graph has a fixed scale from 0–100.

REFERENCES

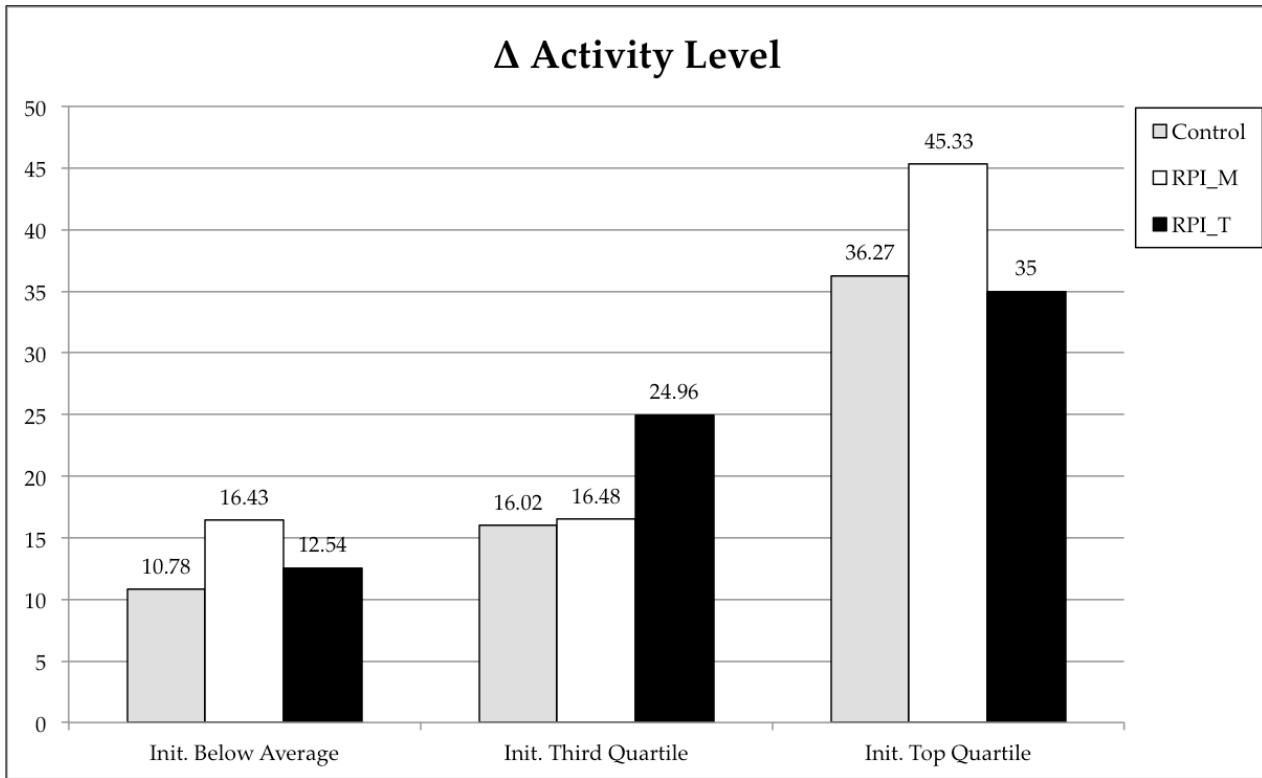
- ABELER, J.; A. FALK; L. GOETTE; AND D. HUFFMAN. "Reference Points and Effort Provision." *American Economic Review* 101(2) (2011): 470–492.
- ALLCOTT, H. "Social Norms and Energy Conservation." *Journal of Public Economics* 95(9–10) (2011): 1082–1095.
- ALLCOTT, H., and T. ROGERS. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104(10) (2014): 3003–3037.
- ALLEN, E.; P. DECHOW; D. POPE; AND G. WU. "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science* 63(6) (2017): 1657–1672.
- ANGRIST, J.; G. IMBENS; AND D. RUBEN. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434) (1996): 444–455.
- ARANDA, C.; J. ARELLANO; AND A. DÁVILA. "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(6 Pt. 1) (1957): 359–372.
- 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.
- BANKER, R. D., AND S. M. DATAR. "Sensitivity, Precision, and Linear Aggregation of Signals for Performance Evaluation." *Journal of Accounting Research* 27(1) (1989): 21–39.
- 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.
- BEBCHUK, L. A., AND J. M. FRIED. "Pay without Performance: Overview of the Issues." *Journal of Applied Corporate Finance* 17(4) (2005): 8–23.
- 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.
- BOL, J. C. "The Determinants and Performance Effects of Managers." *Performance Evaluation Biases.* *The Accounting Review* 86(5) (2011): 1549–1575.
- 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.
- BOUWENS, J., AND P. KROOS. "Target Ratcheting and Effort Reduction." *Journal of Accounting and Economics* 51 (2011): 171–185.
- 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.

- CAMPBELL, D. "Choose the Right Measures, Drive the Right Strategy." *Balanced Scorecard Report* May–June (2006).
- CASAS-ARCE, P., AND F. A. MARTINEZ-JEREZ. "Relative Performance Compensation, Contests, and Dynamic Incentives." *Management Science* 55(8) (2009): 1306–1320.
- CHEN, Y.; F. M. HARPER; J. KONSTAN; AND S. X. LI. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens." *American Economic Review* 100(4) (2010): 1358–1398.
- CROSS, S., AND L. MADSON. "Models of the Self: Self-construals and Gender." *Psychological Bulletin* 12 (1997): 5–37.
- 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.
- EAGLY, A. H. "Sex Differences in Influenceability." *Psychological Bulletin* 85 (1978): 86–116.
- EREZ, M., AND I. ZIDON. "Effects of Goal Acceptance on the Relationship of Goal Difficulty to Performance." *Journal of Applied Psychology* 69 (1984): 69–78.
- 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.
- 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.
- GONG, G.; L. Y. LIN; AND J.Y. SHIN. "Relative Performance Evaluation and Related Peer Groups in Executive Compensation Contracts." *The Accounting Review* 86(3) (2011): 1007–1043.
- 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." *Journal of Public Economics* 148 (2017): 14–31.
- 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.
- HANNAN, R. L.; G. P. MCPHEE; A. NEWMAN; AND I. TAFKOV. "Designing a Performance Feedback System in a Multi-Task Environment: Relative Performance Information Detail Level and Temporal Aggregation in a Multi-Task Environment." Unpublished paper, 2017. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2911072
- HANUSHEK, E. A.; J.F. KAIN; J. M. MARKMAN; AND S. G. RIVKIN. "Does Peer Ability Affect Student Achievement?" *Journal of Applied Econometrics* 18(5) (2003): 527–544.
- HEATH, C; R. LARRICK; AND G. WU. "Goals as Reference Points." *Cognitive Psychology* 38(1) (1999): 79–109.

- HOLMSTROM, B. "Moral Hazard and Observability." *Bell Journal of Economics* (1979): 74-91.
- HOXBY, C. M., AND G. WEINGARTH. "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects." Presented at the 2006 American Economics Association Annual Meetings, 2006. Available at:
www.aeaweb.org/annual_mtg_papers/2006/0108_1300_0803.pdf
- IOANNOU, I; S. X. LI; AND G. SERAFEIM. "The Effect of Target Difficulty on Target Completion: The Case of Reducing Carbon Emissions." *The Accounting Review* 91(5) (2016): 1467–1492.
- KAHNEMAN, D., AND A. TVERSKY. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica* 47(2) (1979): 263–292.
- KASS, R., AND A. RAFTERY. "Bayes Factors." *Journal of the American Statistical Association* 90(430) (1995): 773–795.
- 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.
- 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., 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.
- LATHAM, G. P.; G. SEIJTS; AND D. CRIM. "The Effects of Learning Goal Difficulty Level and Cognitive Ability on Performance." *Canadian Journal of Behavioural Science* 40(4) (2008): 220–229.
- LEONE, A. J., AND S. ROCK. "Empirical Tests of Budget Ratcheting and its Effect on Managers' Discretionary Accrual Choices." *Journal of Accounting and Economics* 33 (2002): 43–67.
- 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 G. P. LATHAM. "Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-year Odyssey." *American Psychologist* 57(9) (2002): 705–717.
- 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.
- MARKLE, A.; G. WU; R. J. WHITE; AND A. M. SACKETT. "Goals as Reference Points in Marathon Running: A Novel Test of Reference Dependence." Unpublished paper, 2015.

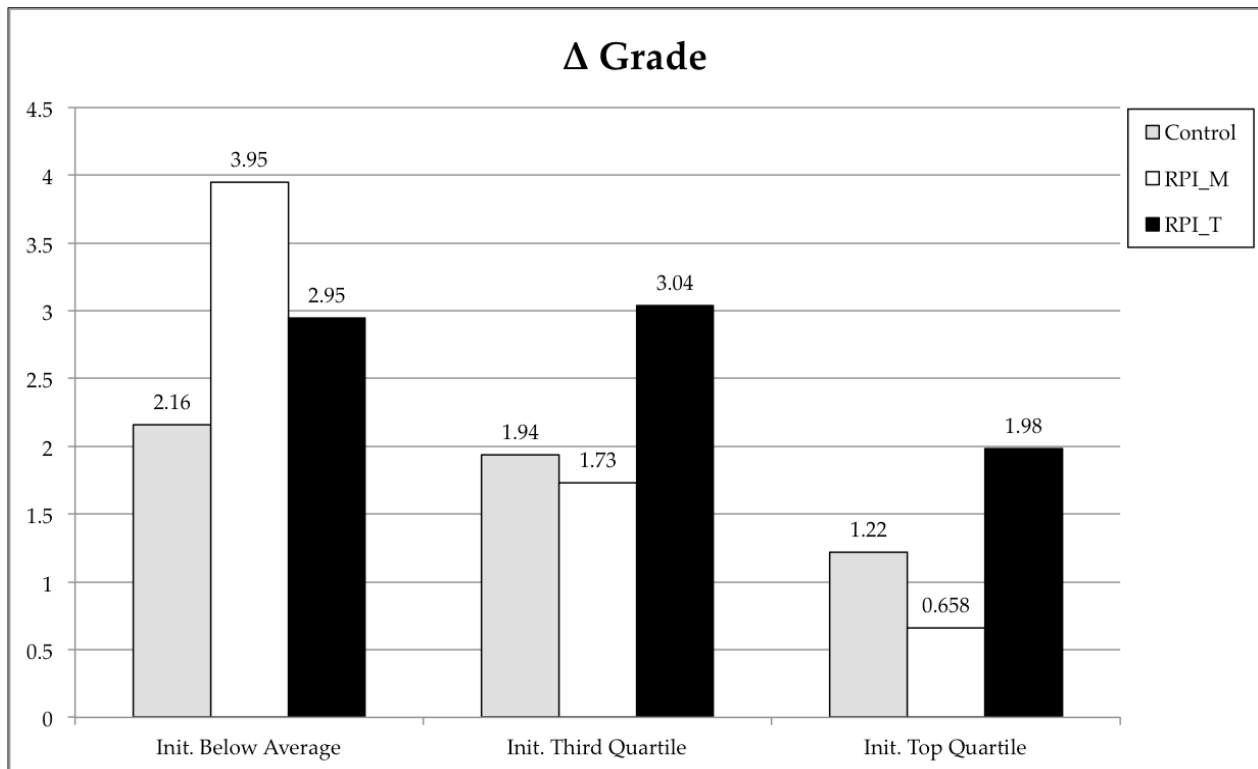
- Available at <http://dx.doi.org/10.2139/ssrn.2523510>
- MERCHANT, K. A., AND J. F. MANZONI. "The Achievability of Budget Targets In Profit Centers: A Field Study." *The Accounting Review* 64(3) (1989).
- MCGREGOR, J. "For Whom the Bell Curve Tolls." *The Washington Post*, 2013. Available at: <https://www.washingtonpost.com/news/on-leadership/wp/2013/11/20/for-whom-the-bell-curve-tolls/>
- MURPHY, K. J. "Performance Standards in Incentive Contracts." *Journal of Accounting and Economics* 30(3) (2000): 245–278.
- 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>
- NEWMAN, A., AND I. TAFKOV. "Relative Performance Information in Tournaments with Different Prize Structures." *Accounting, Organizations and Society* 39(5) (2014): 348–361.
- 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).
- 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.
- 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.
- SMITH, R. (2000). "Assimilative and Contrastive Emotional Reactions to Upward and Downward Social Comparisons." In J. Suls & L. Wheeler, *Handbook of Social Comparison: Theory and Research*. Springer Science & Business Media, 2013.
- SONG, H.; A. TUCKER; K. MURRELL; AND D. VINSON. "Closing the Productivity Gap: Improving Worker Productivity Through Public Relative Performance Feedback and Validation of Best Practices." Unpublished paper, (2017). Available at: <http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2017.2745>
- TAFKOV, I. "Private and Public Relative Performance Information under Different Compensation Contracts." *The Accounting Review* 88(1) (2013): 327–350.
- VANEK SMITH, S. "I, Waiter." *National Public Radio*. (2015) Available at <http://www.npr.org/templates/transcript/transcript.php?storyID=407086723>
- VROOM, V. H. *Work and Motivation*. Wiley, 1964.
- XIAO, Y., AND R. LUCKING. "The Impact of Two Types of Peer Assessment on Students' Performance and Satisfaction within a Wiki Environment". *Internet and Higher Education* 11(3–4) (2008): 186–193.

Figure 1.



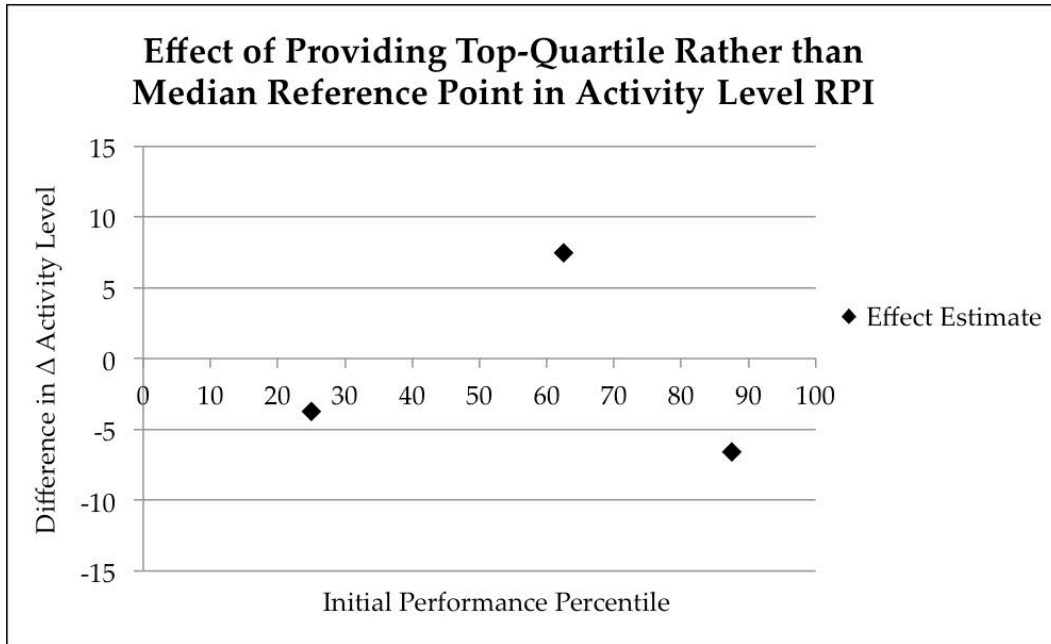
This figure shows the mean Δ Activity Level for each treatment and in each partition of initial performance. These data are from the main experiment, with Activity Level as the displayed measure of performance in RPI.

Figure 2.



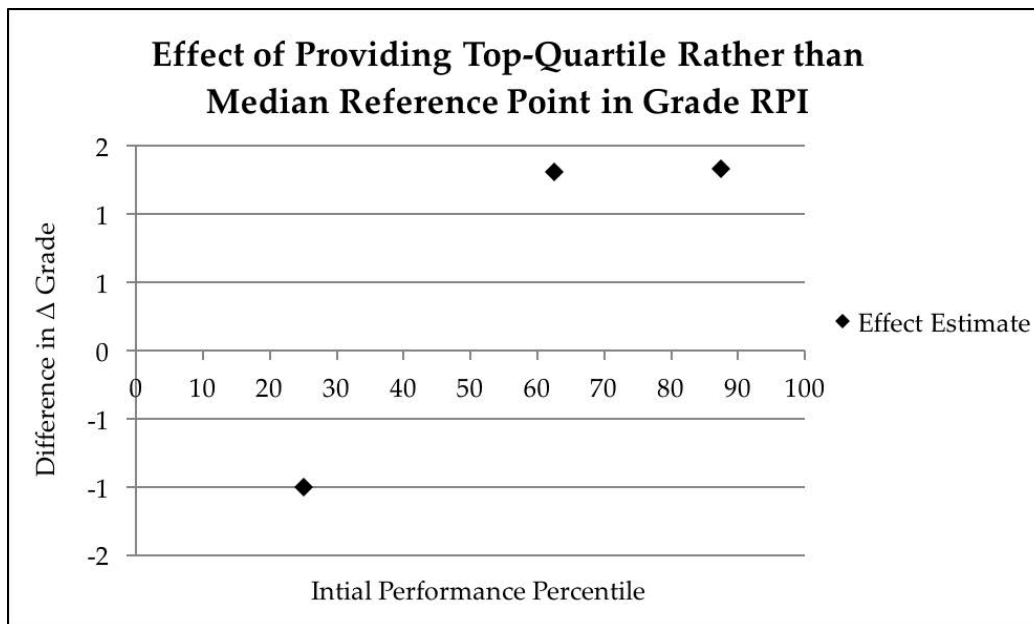
This figure shows the mean Δ Grade for each treatment and in each partition of initial performance. These data are from the supplemental experiment, with Grade as the displayed measure of performance in RPI.

Figure 3.



This figure plots estimated effects on Δ Activity Level of providing the top-quartile rather than median reference point in Activity Level RPI. The effects are estimated in partitions of initial performance (below median, third quartile, and above top quartile) from the main experiment.

Figure 4.



This figure plots estimated effects on Δ Grade of providing the top-quartile rather than median reference point in Grade RPI. The effects are estimated in partitions of initial performance (below median, third quartile, and above top quartile) from the supplemental experiment.

Table 1. Sample Selection and Descriptive Statistics for Main Experiment

Panel A: Sample Selection					
Total Enrollment					24,554
Exclude Students who did not Access the Course					9,383
Final Sample					15,171
Panel B: Descriptive Statistics					
	<u>N</u>	<u>Mean</u>	<u>Std. Dev.</u>	<u>25%</u>	<u>75%</u>
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	11,441	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	15,171	117.73	247.74	5	70
Δ Activity Level	15,171	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 registrations 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. This was a planned analysis.

Table 2. 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

	<u>N</u>	<u>Mean</u>	<u>Std. Dev.</u>	<u>25%</u>	<u>75%</u>
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. This was an application to the supplemental experiment of a planned analysis for the main experiment.

Table 3. RPI Access in Main Experiment

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

This table describes the the distribution of RPI 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. This was a planned analysis.

Table 4. RPI Access in Supplemental Experiment

Sample	N
<i>RPI</i>	
Never Accessed RPI	2,214
Accessed RPI Once	409
Accessed RPI More than Once	372
<i>RPI_M</i>	
Never Accessed RPI	1,120
Accessed RPI Once	203
Accessed RPI More than Once	180
<i>RPI_T</i>	
Never Accessed RPI	1,120
Accessed RPI Once	206
Accessed RPI More than Once	192

This table describes the distribution of RPI access in the supplemental experiment. The table categorizes individuals by experimental condition. The difference in RPI access between the RPI_M and RPI_T groups are not statistically significant. This was an application to the supplemental experiment of a planned analysis for the main experiment.

Table 5. 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 35.00 (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. This model describes cells used in testing planned hypotheses.

Table 6. 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. This model describes cells used in testing planned hypotheses from the main experiment, applied to the supplemental experiment.

Table 7. Hypothesis Tests for Main Experiment

Cells	Coefficient	Intercept	Hypothesis	N	Test Statistic	P-Value
$c_{1...3}$	17.65	None	$H1_0: c_{1...3} = c_{4...9}$	15,171	$F = 5.12^{**}$	0.023
$c_{4...9}$	21.49		$H1_A: c_{1...3} < c_{4...9}$			
c_4	16.43	None	$H2a_0: c_4 = c_7$	5,882	$F = 4.79^{**}$	0.028
c_7	12.54		$H2a_A: c_4 > c_7$			
c_5	16.48	None	$H2b_0: c_5 = c_8$	2,162	$F = 4.84^{**}$	0.027
c_8	24.96		$H2b_A: c_5 < c_8$			
c_6	45.33	None	$H2c_0: c_6 = c_9$	2,288	$F = 2.45$	0.117
c_9	35.00		$H2c_A: c_6 < c_9$			
$c_7 \& 9$	-4.86	$c_4 \& 6$	$H3a_0: c_7 \& 9 = c_8$	10,332	$\chi^2 = 11.13^{***}$	0.000
c_8	7.43	c_5	$H3a_A: c_7 \& 9 < c_8$			
c_7	-3.71	c_4	$H3b_0: c_7 = c_8$	8,044	$\chi^2 = 9.09^{***}$	0.002
c_8	7.43	c_5	$H3b_A: c_7 < c_8$			
c_8	7.43	c_5	$H3c_0: c_8 = c_9$	4,450	$\chi^2 = 7.46^{***}$	0.006
c_9	-6.57	c_6	$H3c_A: c_8 > c_9$			
$c_{4...6}$	14.34	None	$H_0: c_{4...6} = c_{7...9}$	10,332	$F = 1.75$	0.185
$c_{7...9}$	12.20					

This table shows hypothesis tests for the main experiment, with Δ Activity Level as the dependent variable. All tests include fixed effects for the experiment host course. The referenced cells are those from Table 5. The sample for each test consists of the cells compared in the hypothesis, along with the intercept cells, if any. F-tests compare Δ Activity Level in the RPI and control conditions, and in the top-quartile and median conditions. χ^2 -tests compare estimated effects on Δ Activity Level of providing the top-quartile rather than median reference point. *, **, *** denote statistical significance at the .1, .05, and .01 levels, respectively. These models test planned hypotheses and yield effect estimates of the same magnitude that are significant at the same levels as those in OLS models from the approved proposal. We use the above model in the main text to more clearly link cells of treatment and initial performance to hypotheses, tests, and results.

Table 8. Hypothesis Tests for Supplemental Experiment

Cells	Coefficient	Intercept	Hypothesis	<i>N</i>	Test Statistic	P-Value
$c_{1...3}$	1.98	None	$H1_0: c_{1...3} = c_{4...9}$	4,460	$F = 9.36^{***}$	0.002
$c_{4...9}$	2.95		$H1_A: c_{1...3} < c_{4...9}$			
c_4	3.95	None	$H2a_0: c_4 = c_7$	2,055	$F = 2.79^*$	0.095
c_7	2.95		$H2a_A: c_4 > c_7$			
c_5	1.73	None	$H2b_0: c_5 = c_8$	467	$F = 1.98$	0.159
c_8	3.04		$H2b_A: c_5 < c_8$			
c_6	0.65	None	$H2c_0: c_6 = c_9$	473	$F = 4.59^{**}$	0.032
c_9	1.98		$H2c_A: c_6 < c_9$			
$c_7 \& 9$	-0.56	$c_4 \& 6$	$H3a_0: c_7 \& 9 = c_8$	2,995	$\chi^2 = 3.15^*$	0.075
c_8	1.31	c_5	$H3a_A: c_7 \& 9 < c_8$			
c_7	-1.00	c_4	$H3b_0: c_7 = c_8$	2,522	$\chi^2 = 4.36^{**}$	0.036
c_8	1.31	c_5	$H3b_A: c_7 < c_8$			
c_8	1.31	c_5	$H3c_0: c_8 = c_9$	923	$\chi^2 = 0.00$	0.991
c_9	1.32	c_6	$H3c_A: c_8 > c_9$			
$c_{4...6}$	3.09	None	$H_0: c_{4...6} = c_{7...9}$	2,995	$F = 0.37$	0.542
$c_{7...9}$	2.81					

This table shows the hypothesis tests for the main experiment, with Δ Grade as the dependent variable. The referenced cells are those from Table 6. The sample for each test consists of the cells compared in the hypothesis, along with the intercept cells, if any. F-tests compare Δ Grade in the RPI and control conditions, and in the top-quartile and median conditions. χ^2 -tests compare estimated effects on Δ Activity Level of providing the top-quartile rather than median reference point. *, **, *** denote statistical significance at the .1, .05, and .01 levels, respectively. These models test planned hypotheses for the main experiment, applied to the supplemental experiment, and yield effect estimates of the same magnitude that are significant at the same levels as those in OLS models from the approved proposal. We use the above model in the main text to more clearly link cells of treatment and initial performance to hypotheses, tests, and results.

Table 9: 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)
3. 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-value: 0.033
N=57

Importance of reaching reference point

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

Confidence in ability to reach reference point

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

This table shows survey questions and responses regarding individuals' opinions of the 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. This was a planned supplemental analysis.

Table 10: 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)
3. 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 (22)
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-value: 1.000

N=41

Importance of reaching reference point

z-score: 0.70 in favor of median reference point

p-value: 0.479

N=39

Confidence in ability to reach reference point

z-score: 2.49 in favor of median reference point

p-value: 0.012

N=39

This table shows survey questions and responses regarding individuals' opinions of the 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. This was an application to the supplemental experiment of a planned supplemental analysis for the main experiment.