Strategic Manipulation in Peer Performance Evaluation:

Evidence from the Field*

Yifei Huang
Microsoft Research, USA
yhuang.econ@gmail.com

Matthew Shum
California Institute of Technology, USA
mshum@caltech.edu

Xi Wu
Central University of Finance and Economics, China
wuxi@cufe.edu.cn

Jason Zezhong Xiao
Cardiff University, UK
Xiao@cardiff.ac.uk

January 2017

* We thank Ming Hsu, Lawrence Jin, Clive Lennox, Jian Ni, Alejandro Robinson, Jean-Laurent Rosenthal, Thomas Ruchti, Robert Sherman and participants in presentations at Caltech and Zhejiang University for useful comments. Xi Wu thanks the managing partner and the head of the human resource department of the participating audit firm for providing proprietary data and information on performance evaluation that make this study possible.

Corresponding author: Matthew Shum, J Stanley Johnson Professor of Economics, Caltech (mshum@caltech.edu).

Strategic Manipulation in Peer Performance Evaluation: Evidence from the Field

ABSTRACT

This study examines strategic behavior in "360-degree" performance appraisal systems, in which an employee is evaluated by her supervisor as well as her colleagues. Using proprietary data from a mid-sized Chinese public accounting firm, we find that employees manipulate their ratings of peers (i.e., colleagues within the same hierarchical rank of the company). Specifically, they downgrade ratings of their more qualified peers while granting higher ratings to their less qualified peers, compared with evaluations from employees who are not peers. Moreover, this manipulation is mostly done by employees who themselves are less qualified. Altogether, this implies that more-qualified employees "lose" from the 360-degree evaluation scheme, and simulations show that their promotion chances would be (slightly) higher under the traditional "top-down" scheme in which their performance ratings are based only on the their supervisor's appraisal.

Keywords: peer performance evaluation, strategic manipulation, personnel economics, field data

I. INTRODUCTION

Accurate and informative performance evaluation is highly valued in many organizations. It is the basis for implementing incentive plans such as merit pay and for making critical personnel decisions such as promotions (Hunt 1995). Traditionally, performance evaluation uses a "top-down" system in which supervisors assess their subordinates (Jiambalvo 1979). However, since information of a specific employee's performance is dispersed among his/her supervisors, peers, subordinates and even business partners, it is reasonable to ask all the relevant people to participate in performance evaluation.

This is the basic idea of 360-degree feedback in the industry.[1] By the 1990s, 360-degree feedback gained huge popularity, and it was estimated that over one-third of US companies and more than 90% of Fortune 500 firms use some form of 360-degree feedback (Bracken, Timmreck, and Church 2001; Edwards and Ewen 1996). Apart from being now considered a common management practice in the industry (Jackson 2012), 360-degree feedback has also been widely applied in the public sector.[2] Initially it was designed as an evaluation tool to assist employee and managerial development, but it has also been widely used as a tool for managerial decision-making such as promotions and compensation (Maylett 2009). In particular, it has been increasingly used as an important means of performance management (Bracken and Churchill 2013). While traditional management accounting-based performance management tools focus on the "what" dimension of performance (e.g., sales, profit, cash flows, and RoE), the 360-degree feedback system focuses on the "how" dimension (i.e., how the results are obtained) (Bracken

[1] In this paper, we use peer performance evaluation and 360-degree feedback interchangeably.

[2] For example, it has been a part of performance evaluation for all medical doctors required by the UK medical regulator, General Medical Council (Ferguson, Wakeling and Bowie 2014).

and Churchill 2013). The two types can be nicely coupled with each other. For example, the 360-degree feedback system can collect feedback from customers on their satisfaction and employees about their knowledge and learning abilities and such feedback can be naturally used for the customer and learning dimensions of the balanced scorecard (Kaplan and Norton 2001a; 2001b).

While the 360-degree feedback system can have many advantages over traditional top-down performance evaluation, it also brings about new challenges, especially with regard to strategic reporting. For instance, as noted by Jack Welch, former CEO of General Electric, "Like anything driven by peer input, the [360-degree feedback] system is capable of being 'gamed' over the long haul" (Welch and Byrne 2003). When the system is used to determine merit pay or promotion, raters likely face a conflict of interest problem in evaluating their work colleagues, who are also potential competitors for promotions. Either wittingly or unwittingly, personal interest can introduce distortions of facts.

Despite the potential for strategic manipulation or gaming in the 360-degree feedback system, empirical research is scant on such peer performance evaluation behavior in the business field.[3] This paper aims to fill this gap. Using proprietary data from a mid-sized Chinese audit firm which uses a 360-degree performance evaluation system as input into its internal promotion decisions, we measure strategic manipulation in the system, and also examine how the manipulation affects promotion outcomes. To our knowledge, this is among the first to detect strategic reporting in the 360-degree performance appraisal system utilizing field data from an actual business entity.

[3] There have been a few experimental studies on peer performance evaluation using undergraduate students as subjects (e.g., Murphy, Cleveland, Skattebo, and Kinney 2004; Wang, Wong and Kwong 2010).

We find several types of strategic manipulation of the peer evaluation system in our study company. First, we find that employees at the firm tend to inflate their ratings of themselves; overall, however, this has a negligible impact on any employee's overall ratings, which are averaged across all the ratings she received from her colleagues at the firm. Second, we find that employees discriminate against "peers" (i.e., those employees who are within the same hierarchical rank, and hence close competitors for promotions). Specifically, employees tend to denigrate qualified peers who have already passed objective requirements for promotion, while giving generous ratings to peers who have not yet passed.[4] Additionally, we find that strategic reporting among peers is driven by less-qualified employees when they rate their more-qualified peers. This last finding is puzzling and difficult to explain motivationally: as less-qualified raters have little chance of being promoted vis-a-vis qualified peers, there is little benefit from giving lower ratings to qualified peers. One possibility is that the less qualified raters are forward-looking and downgrade their more-qualified peers not to enhance their chances of promotion immediately, but rather in order to reduce *future* performance standards. Alternatively, this finding is also consistent with psychological theories of envy and, to our knowledge, may represent some of the first quantitative evidence of this in a field setting.

Our results imply that more-qualified employees "lose" from the 360-degree evaluation scheme, as their promotion chances would be higher under the traditional "top-down" scheme in which their performance ratings is based only on the appraisal of their superiors. However, simulations using our parameter estimates demonstrate that these differences in promotion probabilities are not economically large. Practically, promotion decisions are based on an

[4] According to the firm's Employee Handbook, the criteria for promotion include both subjective (i.e., 360-degree ratings) and objective (e.g., attendance, academic qualifications, project experience, and tenure) requirements.

employee's *aggregate* rating, which is an average of all the ratings she received from her colleagues at the firm, and this averaging naturally limits the damage that the strategic manipulation by any subset of employees can cause.

Our study makes several contributions to the performance evaluation literature. First, there is an emerging literature identifying potential biases during the performance evaluation process in a traditional top-down evaluation regime (e.g., Bol 2011; Du, Tang, and Young 2012). Our study provides evidence of evaluation biases in an alternative regime, i.e., 360-degree (or peer) performance evaluation. Moreover, we are among the first to provide *field-based* evidence of strategic reporting in the 360-degree feedback system. This complements recent experimental-based literature on tournaments, performance appraisal and sabotage (e.g., Carpenter, Matthews, and Schirm 2010; Harbring and Irlenbusch 2011; Berger, Harbring, and Sliwka 2013). Finally, our study has implications for users of 360-degree appraisal systems to better understand the potential sources of evaluation bias and for improving the practice of peer performance evaluation. For example, to offset the biases identified in this study, decision makers may underweight self-ratings and ratings from less qualified peers. In addition, the decision makers may also consider incorporating disciplinary or incentive measures into 360-degree feedback to dampen strategic behavior, or encourage straightforward peer evaluation.

In Section 2, we review the related literature and bring about our research questions. In Section 3, we describe our data and study company. Section 4 presents our empirical approach and results for detecting strategic manipulation in 360-degree performance ratings. In Section 5, we examine the connection between performance ratings and promotion probabilities in the

study company. In Section 6, we use the results from the preceding sections to conduct counterfactual exercise aimed at showing how much strategic manipulations of peer ratings influence promotion outcomes, and also how outcomes would differ between peer evaluation vs. traditional "top-down" evaluation systems. Section 7 concludes.

II. PRIOR RESEARCH AND RESEARCH QUESTIONS

Literature on Subjective Performance Evaluation

Subjective performance evaluation is pervasive in practice, since oftentimes employees' performance can hardly be captured only using objective measures (Prendergast 1999). There has been a body of theoretical literature on optimal incentive contracting with both objective measures and subjective evaluations (e.g., Baker, Gibbons, and Murphy 1994; Levin 2003; MacLeod 2003).[5] Some recent studies also consider the role of peer evaluation. Kim (2011) investigates how peer evaluation can be used to elicit information from a group of coworkers competing for promotion when the manager only has limited knowledge about performance. Deb, Li, and Mukherjee (2016) study the optimal use of peer evaluation in a relational contract setting. Cheng (2015) studies how the optimal contracting depends on the degree of subjectivity of evaluations.[6] Our field-based empirical study complements these analytical model-based studies of peer evaluations.

[5] Also see a review by Bol (2008).

[6] The level of subjectivity is the extent to which signals received by workers about a particular coworker are correlated. Less correlation means more subjective.

Our field study also shares features with laboratory experiments on tournaments, performance appraisal and sabotage.[7] Carpenter, Matthews, and Schirm (2010) explore sabotage in a real effort experiment, where peer assessment is used to determine the allocation of tournament prize. Among other results, they find that when sabotage is more likely, participants exert less effort recognizing that their performance would not be fairly recognized by their peers. Harbring and Irlenbusch (2011) and Berger, Harbring, and Sliwka (2013) show that although tournament structures (or relative performance schemes in general) have the potential to incentivize higher effort, it also induces higher sabotage, which can reverse the incentive effects.

Subjective performance evaluation has also been a long standing management accounting research topic. However, the focus of the literature has been on the traditional top-down managerial appraisal of subordinates (e.g., Govindarajan 1984; Ittner, Larcker and Meyer 2003); determinants of managers' use of subjectivity in performance evaluation (e.g., Gibbs, Merchant, Van der Stede and Vargus 2004; Rajan and Reichelstein 2006; Bol and Smith 2011; Bol, Kramer and Maas 2016), and the effect of subjective measures on managers' performance evaluation biases (Bushman, Indjejikian and Smith 1996; Moers 2005) and on ratee and organizational performance (e.g., Banker, Potter and Srinivasan 2000). Management accounting research remains largely silent on strategic performance evaluation behavior, although managers' strategic external financial reporting behavior is well-documented (e.g., Roychowdhury 2006; Bowlin 2009; Stubben 2010).

[7] For example, the field setting of using peer performance evaluation in the decision making process of employee promotion is one form of tournaments. Also, we are concerned about whether there exists strategic manipulation (a form of sabotage) on the part of raters who likely face a conflict of interest problem in evaluating their work colleagues.

Nevertheless, Ittner, Larcker and Meyer (2003) find that great subjectivity of balanced scorecard-based bonus plan was perceived to give rise to favoritism in bonus awards in their case company. This implies that managers' appraisal of employees takes into consideration of personal relationships. Bol (2011) argues that both centrality bias and leniency bias in performance evaluation are managers' defensive mechanisms to alleviate ramifications of truthful ratings. From a case study, she finds that both centrality bias and leniency bias are positively affected by information gathering costs and strong employee-manager relationships, but they do not necessarily damage employee performance. Du, Tang and Young (2012) exploit a research context where the Chinese government (as superior) evaluates Chinese state-owned enterprises (as subordinates). They find that a subordinate and a superior engage in both influence activities (bottom-up) and favoritism (top-down) in subjective performance evaluation. However, these studies only investigate managers' strategic performance evaluation behaviors, rather than the strategic behaviors of peers in a 360-degree feedback system which is the focus of our study.

Literature on 360-Degree Feedback

The prior literature has explored various aspects of 360-degree feedback, including purposes and goals of the system, development of the system, implementation of the system, and use and effects of the feedback.[8] A main concern of researchers and practitioners is whether 360-degree feedback works. They find that such systems can indeed improve individual or team

[8] For comprehensive reviews, see Morgeson, Mumford, and Campion (2005), Nowack and Mashisha (2012), and Iqbal, Akbar and Budhwar (2015)..

performance and lead to behavioral change under certain conditions,[9] but the effect sizes are modest and such systems may even result in disengagement and performance deterioration when poorly designed or implemented (Nowack and Mashisha 2012).

Hoffman, Lance, Bynum and Gentry (2010) classify the factors that explain variation in performance, and stress that some degree of inter-rater and inter-rater-group disagreement is useful as their ratings may represent different and useful information (Scullen, Mount, and Goff 2000). Extant studies comparing the ratings of different groups (sources) show that there is typically a weak correlation between self-ratings and the ratings of other groups; in addition, the ratings of peers and supervisors are associated with each other to a greater extent, and supervisors' ratings tend to be most reliable (Nowack and Mashisha 2012). Nowack (2009) documents that supervisors are more likely to focus on performance-related behaviors whereas subordinates stress interpersonal and relationship behaviors. While mixed evidence exists (Sala and Dwight 2002), peers have been found to accurately assess ratees' performance, and at times more so than subordinates and managers (e.g., Inceoglu and Externbrink 2012).

Previous studies have also documented that rater accountability (ie. the rater is required to justify her ratings) enhances rating accuracy (Mero and Motowidlo 1995; Murphy 2008). Other studies show that rater accuracy can vary by demographic variables (Iqbal, Akbar and Budhwar 2015) and also depend on a rater's personal likes or dislikes (Antonioni and Park 2001).

Prior studies conceptualize rating inaccuracy as a result of unintentional errors primarily from two perspectives. First is the psychometric perspective, which sees rating errors as the

[9] For example, Smither, London, and Reilly (2005) identify several success factors that represent some of the conditions under which 360-degree feedback works: goal-setting versus implementation intentions; the delivery and content of the feedback; interpretations and emotional responses to feedback; the participant's personality and feedback orientation; readiness to change; and beliefs about change,self-esteem and self-efficacy.

outcome of the rating stimuli's failure to trigger reliable and valid responses (Cronbach 1955). The other is a cognitive perspective, according to which rating errors arise from the limitations of human cognition, such as memory accessibility, cognitive style, and affect (Robins and DeNisi 1993; DeNisi 1996). However, researchers have recently begun to investigate whether rater intention/goals have an influence on rating accuracy. Murphy, Cleveland, Skattebo, and Kinney (2004) document that student raters with specific goals (e.g., to identify teachers' weaknesses, strengths, to give fair assessment, or to motivate the teachers) give ratings consistent with the goals, and give different ratings conforming to different goals. In an experimental setting, Wang, Wong and Kwong (2010) document instances of strategic inflation or deflation of peer ratings. Our study contributes to this emerging literature by documenting a new form of strategic peer rating in a real business environment.

Research Questions

When 360-degree performance evaluations are used to determine merit pay or promotion, raters are often faced with conflict of interest problems. That is, raters and ratees are potential competitors for limited promotion opportunities, so that the 360-degree appraisal system likely elicits strategic reporting by a rater acting in her personal interest, which can introduce distortions of facts. The rater's strategic reporting arsenal can include inflating self-evaluations and deflating the ratings given to others. Such strategic manipulation aims to benefit the rater and hurt the ratee but, more broadly, distorts the overall accuracy and effectiveness of performance evaluation, and harms the interests of the firm. Therefore, our main research questions are (1) to examine whether raters do, indeed, report strategically when evaluating their

colleagues, and if so, to determine (2) under what circumstances the strategic reporting takes place, and (3) to what extent the manipulation biases appraisal results and promotion outcomes

Since there are few models of strategic behavior in a peer evaluation setting to guide our work, we use a flexible approach to assessing the degree of strategic behavior, and "let the data speak for themselves." With that said, there is still a basic rationale regarding the first research question of looking for strategic behavior. That is, a rater's perceived benefit from strategically downgrading a ratee increases with the rater's perceived degree of competition between the two. Since the benefits from downgrading should be largest vis-a-vis those colleagues with whom a rater is directly competing for a promotion, the extent that the rater downgrades a ratee should depend on the perceived intensity of competition between them. More direct and more intense competition between the rater and ratee leads to more aggressive manipulation.

On the other hand, strategically downgrading peers is not costless. That is, once strategic downgrading is detected, it likely tarnishes the rater's reputation for integrity, and may lead to punishment or revenge. For this reason, a rater will typically not simply downgrade all her colleagues across the board, and may have incentives to mask her intentions by granting inflated ratings to non-rivals.

III. DATA

Background on the Field Study Firm

The data used in our study were retrieved from a Chinese audit firm's personnel archive and performance appraisal archive, covering a five-year period from 2010 to 2014. The participating firm ranks between 10th and 20th during our sample period according to the Chinese Institute of

Certified Public Accountants' national ranking of public accounting firms, and has the license to audit Chinese listed companies as well. The main business lines include audits, asset appraisals, and other accounting services. The audit firm adopts a 13-level hierarchical system for each practicing office, ranging from the partner (level 1) to the intern (level 13). The normal promotion decision involves employees ranging from level 12 (junior audit assistant) to level 2 (department head). As specified in the firm's Employee Handbook, each employee's major financial benefits (such as salaried and performance-based compensation) are directly linked to the rank of position. As a trial, one of the firm's practicing offices has been using the 360-degree approach for employee performance evaluation since the evaluation year 2010. The managing partner of the firm decided to implement the 360-degree approach in our study office to form a more comprehensive basis of performance evaluation.

A 360-degree appraisal procedure was conducted annually in the office for the period from 1 July of year *t* to 30 June of year *t*+1, which serves as the basis for promotion decisions.[10] Within each of the 7 engagement departments of the office, every employee is asked to evaluate everyone else *within* the department as well as to conduct self-evaluation. The human resource (HR) department sends a soft copy of a blank evaluation form (with necessary instructions) shortly after the end of the evaluation year *t* (i.e., early July of year *t*+1) to all formal employees. The maintenance of anonymity of each participant's evaluations is instructed in the evaluation form. The HR department collects filled forms directly from each employee through a specified email address within two weeks, and computes each employee's evaluation outcome. The

[10] For example, the performance evaluation year of 2014 covers the period from July 1, 2014 to June 30, 2015.

evaluation outcomes (either in terms of scores or rankings) are not disclosed to employees except for department heads.

In what follows, we will use the terms "rater" and "ratee" to refer to, respectively, a given employee and one of the colleagues that she is asked to rate. As shown in the Appendix, in the original evaluation forms, a rater needs to evaluate the ratee along four broad dimensions (i.e., general knowledge, technical capabilities, comprehensive capabilities, and team working and management), including 30 detailed items. A 0-to-10 numeric scale is used in each evaluation item, where 0 indicates the poorest performance and 10 the best. In the office's incentive system, only the overall ratings (i.e., averaged over all the 30 items) are used. In our study, we use these aggregated overall ratings.

After the 360-degree performance appraisal, the office managing partners (OMPs) and all department heads meet to discuss promotion decisions. According to the firm's promotion guidelines, there are two requirements that an employee needs to meet in order to be promoted. First, the relative ranking of her performance appraisal rating must be among the top 50% in the group of employees at the same level in her department. Second, for each level, there are some objective qualifications for promotion, including attendance, academic qualifications, project experience, and tenure. The HR department records these qualifications and employees know whether they meet these qualifications or not. Based on our interviews with the office's HR department head, department heads go though the promotion decisions for most employees very quickly as the criteria for promotion are well specified by the firm's Employee Handbook. Most of the discussion takes place when the department head proposes any exceptional

recommendation, as any exceptional decisions made on an employee in one department can be observable to (and thus have implications for) employees in other departments.

Descriptive Statistics of the Dataset

Each observation in our dataset is a rating record, specifying the year of rating, the rater, the ratee, performance rating (averaged over the 30 dimensions), and information about the rater and ratee (e.g., department affiliation, rank at the time of performance evaluation, age, gender, educational background). We have a total of 7,778 rater-ratee-year observations for the five years comprising 153 unique employees in 7 departments of the firm.[11]

Panels A and B of Table 1 presents descriptive statistics of the departmental and pre-evaluation hierarchical distributions of the practicing office averaged over our sample period, respectively. Panel A shows that the number of employees participating in the 360-degree appraisal in each department is relatively small, ranging from 4.4 to 20.4. This shall facilitate the advantage of 360-degree appraisal regarding the knowledge among participants of the appraisal scheme. Panel B shows that there are 12.3% at manager levels and 87.7% at assistant levels in the office. Panel C presents yearly statistics of aggregate overall ratings. The mean overall rating increases from 7.70 in 2010 to 8.73 in 2014, which likely indicates an increasing trend of rating leniency over the sample period. Panel D shows that the likelihood of getting promoted after annual performance evaluation is 53.6% on average, ranging from 37.0% in 2013 and 81.2% in 2011.

[INSERT TABLE 1 HERE]

[11] In the 7,778 total ratings, 432 of them are self-ratings and the rest, 7,346, are non-self-ratings. The observations of self-ratings are not included in our main analysis but are used in some supplementary analyses.

IV. DETECTING STRATEGIC MANIPULATION

Preliminary Evidence of Strategic Behavior

We start by providing some simple evidence showing that employees are indeed exhibiting self-interest in their rating behavior. We ask a simple question: how much do an employee's own ratings (of herself and of others) lead to a better performance ranking in the department than what she actually achieves in the appraisal? In other words, how would an employee's appraisal result be improved if the result was wholly dictated by her own ratings (of herself and of others)? To answer this question, we define a measure $\Delta PR_{self} = PR_{self} - PR_{actual}$, where PR_{self} is the employee's percentile rank[12] according to her own rating and PR_{actual} is her percentile rank in the actual appraisal result. Since higher percentile rank corresponds to better relative ranking, a positive ΔPR_{self} implies that the employee's relative ranking according to her own ratings is better than what she actually achieves in the appraisal. Figure 1 is the histogram of ΔPR_{self} and Table 2 presents the summary statistics of these three variables.

The results suggest that an employee's percentile rank is substantially higher according to her own ratings, compared with the actual appraisal result. Specifically, on average an employee would improve her percentile rank by about 6.3% if the appraisal result was dictated by her own ratings; alternatively, raters systematically rank themselves among the top half of employees in the department, thus placing themselves above the bar (which is at 50%) to satisfy the promotion requirement. At face value, these results suggest that employees do manipulate their self-ratings.

[INSERT FIGURE 1 AND TABLE 2 HERE]

[12] If there are n people and an employee is ranked as the k-th highest, then her percentile rank is (k-1)/(n-1). She gets a percentile rank of 1 if she obtains the highest rating, while a percentile rank of 0 corresponds to the poorest rating in the department.

Ratee Qualification and Strategic Rating

As discussed in section 2.2, our rationale for detecting strategic reporting behavior is that a rater's perceived benefit from strategically downgrading a ratee increases with the rater's perceived degree of competition for a promotion between the two. Therefore, we examine how an employee's rating of a particular colleague depends on variables which are related to whether the rater and the ratee are in a greater degree of competition for a promotion. The two main variables we consider are, first, whether the two employees are "peers", in the sense that they are in the same hierarchical rank within the organization; and, second, whether either of these employees are qualified in that they have already passed objective hurdles for promotion. First, a rater is more likely to compete with her peers (rather than nonpeers) for a promotion, because the promotion is usually made among each rank of employees to its next higher one, with very rare exceptions of a leap across ranks with a gap. Second, more qualified employees could be more likely to get promoted, thus imposing a greater threat to a rater.

Empirical Specification

Our main empirical model is the following:

$$RATING_{ijt} = \beta_0 + \beta_1 PEER_{ijt} + \beta_2 RateeQual_{jt} + \beta_3 PEER_{ijt} \times RateeQual_{jt} + FE_{RateeRank} + FE_{Ratee} + FE_{Dept} + FE_{Year} + \varepsilon_{ijt} \quad (1)$$

In the regression equation (1), $RATING_{ijt}$ denotes the rating that rater i gives to ratee j at year t. The rating scale ranges from zero to ten with 0 as denoting the poorest performance and 10 denoting the highest level of performance. We define the variable $PEER_{ijt}$ equal to one if the rater

i and ratee *j* are of the same rank in year *t*, and zero otherwise. We define the variable $RateeQual_{jt}$ equal to one if the ratee *j* has already passed minimum requirements for promotion (in terms of attendance, academic qualifications, project experience, and tenure) at year *t*, and zero otherwise. We control for fixed-effects specific to each ratee, ratee's rank, ratee's department, and evaluation year.

The interaction terms in the regression model are important for quantifying strategic manipulation. In the absence of strategic manipulation, we expect peer raters and non-peer raters to behave similarly in recognizing the achievements of their peers. Specifically, how much a rater rewards rates who have passed promotion requirements should not depend on whether the ratee and the rater are peers. In the regression model, this implies that the interaction of *PEER* and *RateeQual*, $\beta_3 = 0$.

However, when strategic manipulation is present, a rater's evaluation would depend on whether the ratee is a peer. Since peers who have failed requirements are less threatening to the rater compared with peers who have passed, a strategic rater would give more generous ratings to peer ratees who have failed requirements, denigrating peer ratees who have passed these requirements. In terms of the coefficients in equation (1), we would expect $\beta_3 < 0$ and $\beta_1 > 0$. Therefore, β_3 is a main manipulation measure, since it captures how much a rater strategically downgrades qualified peers who have already passed important promotion requirements. Moreover, β_1 is a derivative "manipulation measure" (for masking), since it captures how much a rater strategically inflates ratings of less qualified peers who have not yet passed these requirements.

Empirical Results

Table 3 presents the estimation results of the OLS regression analysis with robust standard errors clustered by ratee and year. Overall, the empirical results suggest that raters do reward non-peer rates for passing the objective promotion requirements (i.e., $\beta_2 > 0$)., However, the regression coefficient of the interaction term between *PEER* and *RateeQual* (i.e., β_3) is significantly negative, suggesting that when the rater is a peer, she actually "punishes" the rate who has passed promotion requirements (i.e., the total effect is $0.165 - 0.493 < 0$, $p < 0.01$). In other words, *ceteris paribus*, if the ratee passed promotion requirements, compared with the case of not passing, he/she would secure higher ratings from non-peer raters, and get lower ratings from peer raters. In addition, the coefficient on *PEER* (i.e., β_1) is significantly positive, suggesting that less qualified raters give higher ratings to less qualified peers.

These results indicate a "discriminatory generosity" on the part of peer raters, leading them to denigrate the relative ranking of peer ratees who have already passed promotion requirements, while overrating peer ratees who have not yet passed promotion criteria. This is consistent with our notion of strategic reporting, which simultaneously introduces two types of biases: the first aimed against more competent competitors, and the other favoring less eligible peers (possibly as a mask of the first bias).

[INSERT TABLE 3 HERE]

Rater Qualification and Strategic Rating

In the preceding analysis, we detected a form of strategic manipulation whereby raters give generous ratings to peers who have not yet passed promotion requirements, while they become

harsher in rating peers who have passed. To further examine the strategic rating behavior, we now explore the question of how the rater's qualification affects his/her rating decision. That is, do qualified raters, who have already passed promotion requirements, behave differently from less-qualified raters who have not yet passed these requirements? Answering this question is important for understanding which group (qualified or unqualified raters) drives our previous results.

Empirical Specification

Our empirical model for investigating this question is as follows:

$$\begin{aligned} RATING_{ijt} = {} & \beta_0 + \beta_1 PEER_{ijt} + \beta_2 RateeQual_{jt} + \beta_3 PEER_{ijt} \times RateeQual_{jt} \\ & + \beta_4 RaterQual_{it} + \beta_5 PEER_{ijt} \times RaterQual_{it} + \beta_6 RateeQual_{jt} \times RaterQual_{it} \\ & + \beta_7 PEER_{ijt} \times RateeQual_{jt} \times RaterQual_{it} \\ & + FE_{RateeRank} + FE_{Ratee} + FE_{Dept} + FE_{Year} + \varepsilon_{ijt} \end{aligned} \quad (2)$$

When $RaterQual_{it} = 0$, equation (2) reduces to the regression equation as in Section 4.2.1. When $RaterQual_{it} = 1$, equation (2) reduces to:

$$\begin{aligned} RATING_{ijt} = {} & (\beta_0 + \beta_4) + (\beta_1 + \beta_5) PEER_{ijt} + (\beta_2 + \beta_6) RateeQual_{jt} + (\beta_3 + \beta_7) PEER_{ijt} \times RateeQual_{jt} \\ & + FE_{RateeRank} + FE_{Ratee} + FE_{Dept} + FE_{Year} + \varepsilon_{ijt} \end{aligned} \quad (3)$$

We labeled the coefficient of the interaction term between *PEER* and *RateeQual* as manipulation measure. When $RaterQual_{it} = 0$, the manipulation measure is β_3; when $RaterQual_{it} = 1$, then it equals $\beta_3 + \beta_7$. Thus β_7 captures the change in manipulation measure between raters who have and have not already passed the promotion requirements. If $\beta_7 > 0$, it implies that

raters who have not yet passed requirements are more manipulative than those who have passed; if $\beta_7 < 0$, it implies that raters who have passed requirements are more manipulative.

Empirical Results

Table 4 presents the estimation results of the OLS regression analysis with robust standard errors clustered by ratee and year. In Table 4, β_7 is significantly positive, indicating that unqualified raters are more manipulative than qualified raters. Notably, β_3 is significantly negative (−0.379) while $\beta_3 + \beta_7$ is not significantly different from zero (= -0.379 + 0.406 = 0.027). These findings suggest that unqualified raters downgrade qualified peers, while qualified raters do not behave so. On the other hand, β_1 is significantly positive (0.322) while $\beta_1 + \beta_5$ is not significantly different from zero (= 0.322 - 0.331 = -0.009), which suggests that unqualified raters over-inflate unqualified peers, while qualified raters do not behave so.

[INSERT TABLE 4 HERE]

From a strategic standpoint, this result, that strategic manipulation is driven by less-qualified raters who have not yet passed promotion requirements, and directed at qualified ratees who have passed these requirements, is surprising. One may have expected the opposite, as raters who have not yet passed the objective promotion requirements stand little chance of being promoted vis-a-vis peers who have passed, and hence there is little benefit from giving these peers a lower rating.

Thus, it is difficult to explain the motivation of the less-qualified raters in downgrading their more qualified peers. One possible economic explanation may be that the less qualified raters are forward-looking and downgrade their more-qualified peers today not to enhance their chances

of promotion today, but rather in order to reduce *future* performance standards. However, if such future considerations dominate, it would be puzzling why the less qualified raters downgrade their more-qualified peers but not their less qualified peers. In fact, we find the less qualified raters even inflate the ratings of those less-qualified peers.

Beyond rational strategic motives, this result is also broadly consistent with existing theories in the literature on envy. Smith and Kim (2007, pp. 46-50), in their review of the psychological literature, define envy as the unpleasant emotion arising when an individual compares unfavorably with others who enjoy an advantage in a desired domain linked to her self-worth. Similarly, the social psychology literature (e.g., Fiske, Cuddy, and Glick 2007) pinpoints envy as arising in scenarios when an agent faces unfriendly, but highly competent individuals. Specific conditions of the peer evaluation environment in our study firm align with factors which have been pointed out in the literature as conducive to envy. Similarities between the envied and the envying and self-relevance of the comparison domain are necessary to make social comparisons relevant (e.g., Salovey and Rodin 1984; Schaubroeck and Lam 2004). Moreover, the people feeling envy need to believe that the desired advantage cannot be easily obtained (e.g., Testa and Major 1990). In our study company, raters and ratees who are peers are within the same rank in the organizational hierarchy of the company, and share many job responsibilities; moreover, contrary to peer ratees who have passed the objective promotion requirements, it is difficult for unqualified raters to change the status of their failure to pass the minimum criteria by the time of performance evaluation. To our knowledge, then, our findings here constitute some of the first quantitative evidence supporting these theories of envy in a field setting.

V. RATINGS AND PROMOTION DECISIONS

Given the evidence discussed in the preceding sections documenting different aspects of strategic manipulation in the 360-degree appraisal system in our study firm, in the remainder of the paper we quantify how this manipulation affects promotion outcomes within the firm. This sheds light on how much employees can benefit from manipulation; clearly, manipulation is not an end in itself, but rather employees hope to increase their chances at obtaining a promotion by manipulating the ratings they give to others. How large are these benefits from manipulation?

The first step in answering this question is to estimate the effects of the performance ratings on employees' promotion probabilities. To do this, we collected annual promotion outcomes from the firm's personnel archive. Subsequently, we estimate empirical specifications to determine how much good performance ratings and passing objective promotion requirements affected an employee's chances of being promoted within the firm.

Table 5 presents estimation results of our logistic regression models using *PROMOTION* as the dependent variable (coded 1 if a ratee gets promoted after the annual performance evaluation scheme, and 0 otherwise). In these regressions, we use as a regressor the within-department *percentile* of an employee's average performance rating in a given year, rather than the raw numerical performance rating. We do so for two reasons. First, the firm uses relative rankings in performance evaluation to specify the minimum requirement for being considered for promotion. Second, the rating percentile provides a more comparable measure across years, since it is invariant to fluctuations of rating leniency over the five years of our sample. In addition, we include in the model $LICENSE_{jt}$ (coded 1 if the ratee j has obtained the

CPA license in year *t*, and 0 otherwise) to control for differences in professional qualifications across ratees. Finally, we include fixed effects for year, department, and (pre-evaluation) rank.

In Table 5, Column 1 presents the specification which includes the percentile of performance rating (*PR_dep*), *RateeQual* and the interaction term between them. The coefficients on *PR_dep* and *RateeQual* are both positive and significant at 1%. The interaction term is negative and significant at 5%. Column 2 presents the specification without the interaction term. While the coefficient on *RateeQual* remains positive, the statistical significance weakened (p-value = 0.13). In both specifications, the *LICENSE* dummy is positive and significant at 1%.

These results suggest, first, that passing the objective requirements contributes to promotion. Second, the negative coefficient on the interaction term indicates that the marginal importance of performance rating decreases as the employee passes promotion requirements. In other words, a good performance rating is more important for those who have not yet passed promotion requirements. These results suggest substitutability between performance rating and passing promotion requirements.[13]

[INSERT TABLE 5 HERE]

VI. POLICY IMPLICATIONS: 360-DEGREE APPRAISAL VS. ALTERNATIVE PERFORMANCE RATING SYSTEMS

In previous rating-level analysis, we identify patterns of strategic manipulation when employees rate their peers. Specifically, we find that employees who had not yet passed

[13] To the extent that raters perceive the pattern of substitutability between meeting promotion requirements and 360-degree ratings, results in Table 5 offer one possible explanation to the fact that raters who have passed promotion requirement are less manipulative (as shown in Table 4), although an alternative one could be that qualified raters have a greater level of integrity.

promotion requirements downgraded their peers who had passed and upgraded their peers who have not yet passed, compared with nonpeer employees' rating behavior. We also find that employees who had already passed promotion requirements did not exhibit this discriminatory behavior. Logically, we expect this to distort the aggregated appraisal results in a direction that benefits those who have not yet passed promotion requirements, who manipulated ratings to improve their relative ranking among their peers. Whether and to what extent the strategic manipulation biases appraisal results and promotion outcomes is a question of significant practical implications.

Another important question is how the results of the 360-degree appraisal differ from that of the traditional "top-down" appraisal system where only supervisors evaluate their subordinates. We examine this question by using department heads' ratings to proxy for counterfactual ratings under the top-down appraisal system. This is reasonable because department heads typically do not face direct competition from their subordinates and the anonymity of department heads ratings is strictly preserved in the audit firm under our study.

In this section, we will explore these two questions. We start by analyzing the correlations between different components of 360-degree performance appraisal, including ratings from department heads, peers, nonpeers, and self-evaluations. Then, based on the historical relationship between appraisal results and promotion records, we examine how promotion outcomes would change if only ratings of one of these components (i.e, department heads, peers, nonpeers, or self-evaluations) are used as the basis for making promotion decisions.

Correlations between Ratings from Department Heads, Peers, Nonpeers, and Self-evaluations

In the 360-degree appraisal system, each employee receives evaluations from his/her department head, peers, nonpeers and also conducts self-evaluation. Do these different components of the aggregate rating agree with each other? To answer this question, we consider correlation patterns between the different components. We aggregate performance ratings at the individual level and, for each employee, we compute his/her overall average rating (*rating_avg*),[14] average rating from the department head (*rating_head*), average rating from peers (*rating_peer*), average rating from nonpeers (*rating_nonpeer*),[15] and average self-rating (*rating_self*). Table 6 presents the correlation matrix of these variables.

Results in Table 6 indicate that the department head's ratings (*rating_head*) are less correlated to ratings from peers (*rating_peer*) than to ratings from nonpeers (*rating_nonpeer*) (0.520 vs. 0.827). Interpreting department heads' ratings as a nonstrategic benchmark, this is consistent with our basic notion that peers are more likely to manipulate their ratings strategically than nonpeers. In addition, department heads' ratings and the overall average ratings are highly correlated (with a correlation coefficient of 0.834). Lastly, average ratings from nonpeers and the overall average ratings have a correlation coefficient as high as 0.982. This suggests that the peer evaluation part of the appraisal only leads to a very limited degree of discrepancy between average nonpeer ratings and the overall ratings in our study. These results remain robust if we use within-department percentiles of average ratings instead of the raw value of average ratings.

[14] We excluded self-evaluations from computing the overall average rating. However, the results remain qualitatively unchanged when self-evaluations are included.

[15] Department heads' ratings are included in computing the average rating from nonpeers. Results are qualitatively similar if we exclude these ratings.

[INSERT TABLE 6 HERE]

Alternative Scenarios

In this section we use our previous results to answer several policy questions of interest. First, how much does the strategic manipulation in peer evaluation which we have uncovered so far affect promotion outcomes? Who are the winners and losers from strategic manipulation? Second, how do the outcomes from 360-degree appraisal differ from the outcomes from the traditional top-down approach, where only supervisors evaluate their subordinates? Who are the winners and losers in moving from the traditional appraisal system to the 360-degree system?

An initial step in answering these questions is to link promotion decisions with appraisal results, so that we can analyze how changes in the latter would affect the former. While recognizing the general challenges involved, we use the empirical relationship between appraisal results and promotion, as estimated in Table 5 (column 1), as the basis for these counterfactual evaluations.

Four Counterfactual Scenarios

There are four counterfactual scenarios to consider.

(1) The scenario where appraisal results are determined only by the department head's ratings (denoted as CS_{head}). This proxies for the rating that an employee would have received in a hypothetical top-down performance evaluation scenario.

(2) The scenario where appraisal results are determined only by the peer evaluation part of the 360-degree appraisal (denoted as CS_{peer}).

(3) The scenario where appraisal results are determined only by the nonpeer part of the 360-degree appraisal (denoted as $CS_{nonpeer}$). That is, all peer evaluations in the original appraisal are dropped. $CS_{nonpeer}$ helps us to see to what extent the peer evaluation part distorts overall promotion outcomes.

(4) The scenario where appraisal results are determined only by self-evaluations (denoted as CS_{self}). Specifically, in this scenario, an employee's percentile rank is determined by the relative ranking of her self-evaluation compared with the ratings she gives to others in the department.[16]

In each case, we will use the corresponding counterfactual appraisal results to compute the counterfactual probability of promotion (denoted as P_{CS}) for each employee, employing the fitted logistic model of promotion decision in Table 5. Then, we compare the counterfactual promotion probabilities with actual promotion probabilities (denoted as P_{actual}) which are the predicted promotion probabilities using the actual appraisal results. We define $\Delta_{CS} = P_{CS} - P_{actual}$ as the increase of promotion probability in the counterfactual scenario compared with the actual case. For example, if an employee's P_{actual} is 0.3 and P_{head} is 0.4, then her Δ_{head} is 0.1. That is, her probability of promotion would increase by 0.1 if the promotion decision was solely determined by her department head's rating.

Results

Table 7 presents the summary statistics of the counterfactual changes in promotion probabilities corresponding to the four scenarios. The most notable result in Table 7 is that employees increased their promotion probabilities if the appraisal results were determined by

[16] If there are n people in her department and she rates herself the k-th highest in the department, then her percentile rank in CS_{self} is (k-1)/(n-1).

their own ratings (cf. the counterfactual scenario of CS_{self}). In this scenario, 42.79% of employees increased their promotion probabilities while 31.84% of employees had decreased promotion probabilities. On average, an employee increased her promotion probability by 3.98% under the counterfactual scenario of CSself, which represents an increase of 7.16% (the average probability of actual promotion is 55.59% in this sample). This result characterizes the extent to which one's self-evaluation promotes self-interest. In fact, this result is driven by that self-evaluation tends to inflate one's own percentile rank. As shown in Table 2, the percentile rank according to one's own ratings (*PR_self*) has a mean of 0.573, while the mean percentile rank in the actual appraisal result is 0.508.

[INSERT TABLE 7 HERE]

Table 8 presents the correlation matrix of these four counterfactual changes in promotion probabilities together with *RateeQual*, the dummy of passing objective promotion requirements. First, Δ_{peer} and $\Delta_{nonpeer}$ have a significantly negative correlation coefficient of −0.283. This is the only pair of Δ_{CS} variables that are negatively correlated, suggesting that these two counterfactual scenarios lead to different consequences upon promotion outcomes. This result is consistent with our expectation that peer evaluation and nonpeer evaluation reflect different motives in the rater. Second, Δ_{peer} and *RateeQual* are negatively correlated (−0.103, and significant at 5%), which is consistent with our earlier finding that ratings by peers are most biased when the rater has not yet passed promotion requirements.

[INSERT TABLE 8 HERE]

To further explore how employees who have and have not yet passed promotion requirements would be differentially affected, we run regressions of counterfactual promotion

probabilities (i.e, Δ_{CS}) on *RateeQual* as well as other determinants of promotion as controlled for in Table 5. Table 9 presents the regression results.

First, Column (1) shows that the coefficient on *RateeQual* is positively and marginally significant at 10% level, which suggests that employees who passed promotion thresholds would have enjoyed a higher promotion probability under the top-down appraisal scenario (CS_{head}), in which performance ratings are determined solely by the department head. This suggests that relatively qualified employees – those who have already passed promotion requirements – could be hurt (although in a small magnitude as suggested in the coefficient) in moving from the traditional top-down appraisal to 360-degree appraisal, due to strategic manipulation.

Second, results in Column (2) show that a less-qualified employee who has not yet passed promotion requirements would be better off under CS_{peer}, as her promotion probability increases by 3.2% in this counterfactual scenario. Therefore, peer evaluation, in the aggregate, benefits those "manipulators" who are the relatively less qualified employees – those who have not yet passed promotion requirements, and echoes our earlier empirical results showing how less qualified raters tend to denigrate their more qualified peers in the rating process. Additionally, the coefficient on *LICENSE* suggests that an employee with the CPA license also would have experienced a decrease in promotion probability by 4.2% in this scenario which incorporates ratings from peers.

Third, Column (3) shows that the coefficient on *RateeQual* is not significantly different from zero under $CS_{nonpeer}$, suggesting that the peer evaluation does not impose an impact substantial enough to benefit those who have not yet passed promotion requirements. One possible reason is that ratings by peers, which are the "problematic" ones, only constitute a small proportion - only

15.14% - of the total ratings within the firm. Moreover, the averaging of individual ratings as part of the 360-degree appraisal system further dilutes the effects of these peer ratings on overall ratings.

Finally, results in Column (4) suggests that self-evaluations do not lead to a discriminatory consequence for those who passed promotion requirements, although we have already shown that, of course, raters would benefit if their overall rating were completely determined their own self-evaluations.

[INSERT TABLE 9 HERE]

VII. CONCLUSION

In this paper we have utilized unique proprietary data from a mid-sized Chinese audit firm to examine the extent of strategic maneuvering in a 360-degree performance evaluation system. Perhaps not surprisingly, we find that employees at the firm tend to inflate their own "self-ratings", but overall this has a negligible impact on any employee's overall ratings, which are averaged across all the ratings she received from her colleagues at the firm. More subtly, we find that employees use different rating schemes to evaluate peers who are within the same hierarchical rank (and hence close competitors for promotions). Specifically, more qualified peers are systematically downgraded, and this effect is driven by less-qualified employees when they rate more-qualified peers. Since less-qualified employees have little obvious gain from downgrading more qualified peers, such behavior is broadly in line with psychological notions of "envy" or "spite".

Counterfactual simulations show that this strategic manipulation of ratings towards qualified peers leads to slightly lower promotion probabilities for these qualified individuals. Moreover, in the scenario where promotional decisions are based only on department heads' ratings of employees (the traditional "top-down" appraisal system), qualified employees would also be promoted with a higher probability than under 360-degree appraisal. On one hand, these changes in promotion probabilities are not economically large, possibly due to any strategic reporting biases being offset by all participants' ratings. On the other hand, they do highlight potential sources of bias inherited in the 360-degree appraisal system, and hence avenues for improvement in the design of such systems.

REFERENCES

Antonioni, D., and H. Park. 2001. The Relationship Between Rater Affect and Three Sources of 360-Degree Feedback Ratings. *Journal of Management* 27: 479-495.

Baker, G., R. Gibbons, and K. J. Murphy. 1994. Subjective Performance Measures in Optimal Incentive Contracts. *The Quarterly Journal of Economics* 109 (4): 1125-1156.

Banker, R. D., G. Potter, and D. Srinivasan. 2000. An Empirical Investigation of an Incentive Plan That Includes Nonfinancial Performance Measures. *The Accounting Review* 75 (1): 65-92.

Berger, J., C. Harbring, and D. Sliwka. 2013. Performance Appraisals and the Impact of Forced Distribution - An Experimental Investigation. *Management Science* 59 (1): 54-68.

Bol, J. 2008. Subjectivity in Compensation Contracting. *Journal of Accounting Literature* 27: 1-24.

Bol, J. 2011. The Determinants and Performance Effects of Managers' Performance Evaluation Biases. *The Accounting Review* 86 (5): 1549-1575.

Bol, J. C., S. Kramer, and V. S. Maas. 2016. How control system design affects performance evaluation compression: The role of information accuracy and outcome transparency. *Accounting, Organizations and Society* 51: 64-73.

Bol, J., and S. D. Smith. 2011. Spillover Effects in Subjective Performance Evaluation: Bias and the Asymmetric Influence of Controllability. *The Accounting Review* 86 (4): 1213-1230.

Bowlin, K. 2009. Experimental Evidence of How Prior Experience as an Auditor Influences Managers' Strategic Reporting Decisions. *Review of Accounting Studies* 14 (1): 63-87.

Bracken, D., and A. Churchill. 2013. The "new" performance management paradigm: capitalizing on the unrealized potential of 360 degree feedback. *People and Strategy* 36 (2): 34-40.

Bracken, D. W., C. W. Timmreck, and A. H. Church. 2001. *The Handbook of Multisource Feedback*. John Wiley & Sons.

Bushman, R. M., R. J. Indjejikian, and A. Smith. 1996. CEO compensation: The role of individual performance evaluation. *Journal of Accounting and Economics* 21 (2): 161-193.

Carpenter, J., P. H. Matthews, and J. Schirm. 2010. Tournaments and Office Politics: Evidence from a Real Effort Experiment. *The American Economic Review* 100 (1): 504-517.

Cheng, C. 2015. Moral Hazard in Teams with Subjective Evaluations. Working paper. Northwestern University.

Cronbach, L. J. 1955. Processes affecting scores on understanding of others and assumed similarity. *Psychological Bulletin* 52: 177-193.

Deb, J., J. Li, and A. Mukherjee. 2016. Relational Contracts with Subjective Peer Evaluations. *The RAND Journal of Economics* 47 (1): 3-28.

DeNisi, A. S. 1996. *A cognitive approach to performance appraisal: A program of research*. New York, NY: Routledge.

Du, F., G. Tang, and S. M. Young. 2012. Influence Activities and Favoritism in Subjective Performance Evaluation: Evidence from Chinese State-Owned Enterprises. *The Accounting Review* 87 (5): 1555-1588.

Edwards, M. R., and A. J. Ewen. 1996. *360 Feedback: The Powerful New Model for Employee Assessment & Performance Improvement*. Amacom.

Ferguson, J., J. Wakeling, and P. Bowie. 2014. Factors influencing the effectiveness of multisource feedback in improving the professional practice of medical doctors: a systematic review. *BMC Medical Education* 14: 76. (http://www.biomedcentral.com/1472-6920/14/76)

Fiske, S. T., A. J. C. Cuddy, and P. Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences* 11 (2): 77-83.

Gibbs, M., K. A. Merchant, W. A. Van der Stede, and M. E. Vargus. 2004. Determinants and effects of subjectivity in incentives. *The Accounting Review* 79 (2): 409-436.

Govindarajan, V. 1984. Appropriateness of accounting data in performance evaluation: An empirical examination of environmental uncertainty as an intervening variable. *Accounting, Organizations and Society* 9 (2): 125-135.

Harbring, C., and B. Irlenbusch. 2011. Sabotage in Tournaments: Evidence from a Laboratory Experiment. *Management Science* 57 (4): 611-627.

Hoffman, B., C. Lance, B. Bynum, and W. Gentry. 2010. Rater source effects are alive and well after all. *Personnel Psychology* 63: 119-151.

Hunt, S. 1995. A Review and Synthesis of Research in Performance Evaluation in Public Accounting. *Journal of Accounting Literature* 14: 107-139.

Iqbal, M., S. Akbar, and P. Budhwar. 2015. Effectiveness of performance appraisal: An integrated framework. *International Journal of Management Review* 17: 510-533.

Inceoglu, I., and K. Externbrink. 2012. Leadership development: Who knows best how well the highflyers perform? Paper presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Ittner, C.D., D.F. Larcker, and M.W. Meyer. 2003. Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard. *The Accounting Review* 78 (3): 725-758.

Jackson, E. 2012. The 7 Reasons Why 360 Degree Feedback Programs Fail. Forbes, (http://onforb.es/Rnu07q)

Jiambalvo, J. 1979. Performance Evaluation and Directed Job Effort: Model Development and Analysis in a CPA Firm Setting. *Journal of Accounting Research* 17 (2): 436-455.

Kaplan, R. S., and D. P. Norton. 2001a. Transforming the balanced scorecard from performance measurement to strategic management: Part I. *Accounting Horizons* 15 (1) 87–104.

Kaplan, R. S., and D. P. Norton. 2001b. Transforming the balanced scorecard from performance measurement to strategic management: Part II. *Accounting Horizons* 15(2): 147-160.

Kim, J.-H. 2011. Peer Performance Evaluation: Information Aggregation Approach. *Journal of Economics & Management Strategy* 20 (2): 565-587.

Levin, J. 2003. Relational Incentive Contracts. *The American Economic Review* 93 (3): 835-857.

Lipe, M. G., and S. E. Salterio. 2000. The balanced scorecard: Judgmental effects of common and unique performance measures. *The Accounting Review* 75 (3): 283-298.

MacLeod, W. B. 2003. Optimal Contracting with Subjective Evaluation. *The American Economic Review* 93 (1): 216-240.

Maylett, T. 2009. 360 Degree Feedback Revisited The Transition From Development to Appraisal. *Compensation and Benefits Review* 41 (5): 52-59.

Mero, N. P., and S. J. Motowidlo. 1995. Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology* 80 (4): 517−524.

Moers, F. 2005. Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Accounting, Organizations and Society* 30 (1): 67–80.

Morgan, A., K. Cannan, and J. Cullinane. 2005. 360° Feedback: A Critical Enquiry. *Personnel Review* 34 (6): 663-680.

Morgeson, F. P., T. V. Mumford, and M. A. Campion. 2005. Coming full circle: Using research and practice to address 27 questions about 360-degree feedback programs. *Consulting Psychology Journal: Practice and Research* 57 (3): 196–209.

Murphy, K. R. 2008. Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Science and Practice* 1 (2): 148−160.

Murphy, K., J. Cleveland, A. Skattebo, and T. Kinney. 2004. Raters who pursue different goals give different ratings. *Journal of Applied Psychology* 89 (1): 158-164.

Nowack, K. 2009. Leveraging multirater feedback to facilitate successful behavioral change. *Consulting Psychology Journal: Practice and Research* 61 (4): 280–297.

Nowack, K., and S. Mashisha. 2012. Evidence-based answers to 15 questions about leveraging 360-degree feedback. *Consulting Psychology Journal: Practice and Research* 64 (3): 157-182.

Prendergast, C. 1999. The Provision of Incentives in Firms. *Journal of Economic Literature* 37 (1): 7-63.

Rajan, M. V., and S. Reichelstein. 2006. Subjective Performance Indicators and Discretionary Bonus Pools. *Journal of Accounting Research* 44 (3): 585-618.

Robbins, T.L., and A. S. DeNisi. 1993. Moderators of sex bias in the performance appraisal process: a cognitive analysis. *Journal of Management* 19: 113–126.

Roychowdhury, S. 2006. Earnings management through real activities manipulation. *Journal of Accounting and Economics* 42 (3): 335-370.

Sala, F., and S. Dwight. 2002. Predicting executive performance with multi-rater surveys: Whom you ask makes a difference. *Consulting Psychology Journal: Practice and Research* 54: 166–172.

Salovey, P., and J. Rodin. 1984. Some Antecedents and Consequences of Social-comparison Jealousy. *Journal of Personality and Social Psychology* 47 (4): 780-792.

Schaubroeck, J., and S. S. K. Lam. 2004. Comparing Lots Before and After: Promotion Rejectees' Invidious Reactions to Promotees. *Organizational Behavior and Human Decision Processes* 94 (1): 33-47.

Scullen, S. E., M. K. Mount, and M. Goff. 2000. Understanding the latent structure of job performance ratings. *Journal of Applied Psychology* 85 (6): 956-970.

Smith, R. H., and S. H. Kim. 2007. Comprehending Envy. *Psychological Bulletin* 133 (1): 46-64.

Smither, J., M. London, and R. Reilly. 2005. Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology* 58: 33-66.

Stubben, S. R. 2010. Discretionary Revenues as a Measure of Earnings Management. *The Accounting Review* 85 (2): 695-717.

Testa, M., and B. Major. 1990. The Impact of Social Comparisons after Failure: The Moderating Effects of Perceived Control. *Basic and Applied Social Psychology* 11 (2): 205-218.

Wang, X. M., K. F. E. Wong, and J. Y. Y. Kwong. 2010. The roles of rater goals and ratee performance levels in the distortion of performance ratings. *Journal of Applied Psychology* 95 (3): 546–561.

Welch, J., and J. A. Byrne. 2003. *Jack: Straight from the Gut*. Grand Central Publishing.

APPENDIX

The Instrument of 360-Degree Performance Evaluation for Each Employee of the Study Firm

	Names of ratees:	A	B	C	…
	Current vocational ranks of ratees:	Senior manager	Medium manager	Senior II assistant	…
Dimension 1: General knowledge	(1) Accounting	X	X	X	X
	(2) Taxation	X	X	X	X
	(3) Auditing	X	X	X	X
	(4) Asset appraisal	X	X	X	X
	(5) Computer	X	X	X	X
	(6) Foreign language	X	X	X	X
Dimension 2: Technical capabilities	(7) Analytical	X	X	X	X
	(8) Adaptability	X	X	X	X
	(9) Writing	X	X	X	X
	(10) Documentation	X	X	X	X
	(11) Oral reporting	X	X	X	X
	(12) Accuracy	X	X	X	X
	(13) Clarity of expression	X	X	X	X
Dimension 3: Comprehensive capabilities	(14) Working attitude	X	X	X	X
	(15) Working initiatives	X	X	X	X
	(16) Innovativeness	X	X	X	X
	(17) Executing ability	X	X	X	X
	(18) Working efficiency	X	X	X	X
	(19) Rule-abiding	X	X	X	X
	(20) Accountability	X	X	X	X
Dimension 4: Team working and management	(21) Project planning	X	X	X	X
	(22) Ability of tutoring	X	X	X	X
	(23) Project scheduling and control	X	X	X	X
	(24) Communication with clients and employees	X	X	X	X
	(25) Financial management	X	X	X	X
	(26) Leadership	X	X	X	X
	(27) Ability to summarize the work	X	X	X	X
	(28) Maintenance with clients	X	X	X	X
	(29) Development of business	X	X	X	X
	(30) Vocational bearing	X	X	X	X

Note: X ranges from 0 to 10 points. As a guidance for raters, the instrument indicates that: a score ranging from 0 to 3 points means "very poor". A score ranging from 3 to 5 points means "poor". A score ranging from 5 to 6 points means "just meet the floor standard". A score ranging from 6 to 8 points means "good". A score ranging from 8 to 10 points means "excellent".

Figure 1

Histogram of ΔPR_{self}

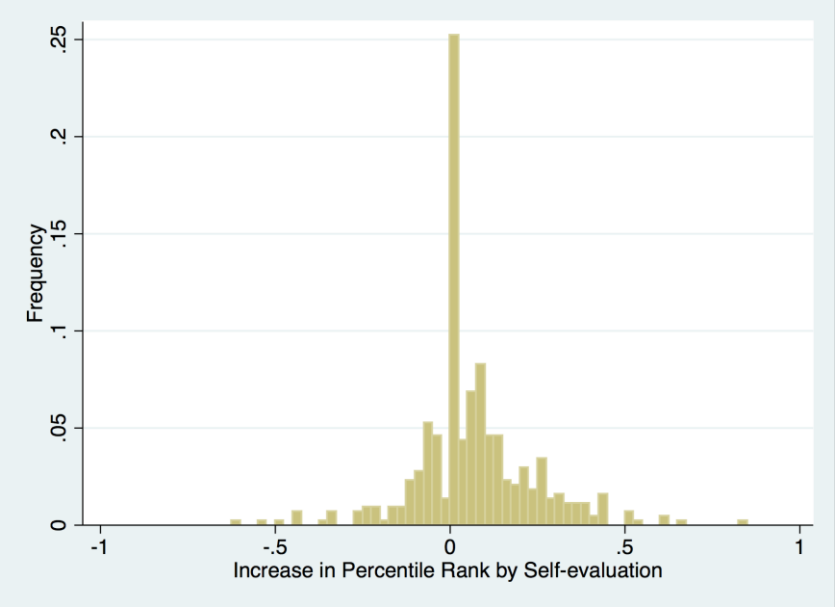

Table 1

Descriptive Statistics for the Study Audit Office

Panel A: Departmental Distribution

Department	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**Total**
Avg. num employees	19.8	11.8	6.8	4.4	19.8	21.6	6.8	91

Panel B: Pre-evaluation Hierarchical Distribution

Pre-evaluation hierarchical level	Avg. num. employees	Percentage
Manager-level	3.8	4.2%
L2: Senior manager	3.2	3.5%
L3: Medium manager	0.2	0.2%
L4: Junior manager	1.0	1.1%
L5: Senior project manager	1.6	1.8%
L6: Medium project manager	1.4	1.5%
L7: Junior project manager	11.2	12.3%
Assistant-level		
L8: Senior II assistant	18.6	20.4%
L9: Senior I assistant	9.0	9.9%
L10: Medium II assistant	15.0	16.5%
L11: Medium I assistant	20.2	22.2%
L12: Junior assistant	17.0	18.7%
Subtotal	79.8	87.7%
Total	91.0	100.0%

Panel C: Annual Statistics of Aggregate Overall Ratings

Evaluation year	N	Mean	S.D.	P10	P25	P50	P75	P90
2010	1,136	7.70	1.20	6.18	7.00	7.72	8.71	9.22
2011	1,478	7.76	1.31	6.01	6.97	7.95	8.77	9.47
2012	1,800	8.20	1.19	6.68	7.42	8.37	9.11	9.58
2013	1,940	8.22	1.17	6.63	7.53	8.43	9.10	9.55
2014	1,424	8.73	1.17	7.00	8.02	9.08	9.60	9.87
2010-2014	7,778	8.14	1.25	6.45	7.30	8.29	9.14	9.62

Notes: In the 7,778 total ratings, 432 of them are self-ratings and the rest 7,346 are non-self-ratings.

Panel D: Promotion Rates

Evaluation year	Num. employees	# Getting promoted	Percentage
2010	72	28	38.9%
2011	85	69	81.2%
2012	95	54	56.8%
2013	108	40	37.0%
2014	95	53	55.8%
2010-2014	455	244	53.6%

Table 2

Summary Statistics of ΔPR_{self}, PR_{self}, and PR_{actual}

Variables	# Obs.	Mean	SD	Min	Q1	Median	Q3	Max
ΔPR_{self}	432	0.065	0.171	-0.625	0.000	0.037	0.140	0.847
PR_{self}	432	0.573	0.300	0.000	0.338	0.611	0.833	1.000
PR_{actual}	432	0.508	0.311	0.000	0.243	0.500	0.778	1.000

Notes: These summary statistics are computed over common observations.

Definition of variables:

ΔPR_{self} = PR_{self} - PR_{actual}.

PR_{self} = the employee's percentile rank according to her own rating.

PR_{actual} = the employee's percentile rank in the actual appraisal result.

Table 3

Ratee Qualification and Performance Rating: Difference between Peer and Nonpeer Raters

Dep. Var.: $RATING_{ijt}$		OLS
$PEER_{ijt}$	β_1	0.436***
		(0.061)
$RateeQual_{jt}$	β_2	0.165***
		(0.042)
$PEER_{ijt} \times RateeQual_{jt}$	β_3	-0.493***
		(0.079)
Ratee rank fixed-effects		Yes
Ratee fixed-effects		Yes
Year fixed-effects		Yes
Department fixed-effects		Yes
N		7,346
Adj. R^2		0.393

Notes: Robust standard errors clustered by ratee and year are reported in parentheses. *, **, *** are significant at 10%, 5%, and 1%, respectively (two-tailed).

Definition of variables:

$RATING_{ijt}$ = the rating that rater i gives to ratee j at year t, ranging from zero (poorest performance) to ten (highest performance).

$PEER_{ijt}$ = 1 if the rater i and ratee j are of the same rank at year t, and 0 otherwise.

$RateeQual_{jt}$ = 1 if the ratee j has already passed minimum requirements for promotion (in terms of attendance, academic qualifications, project experience, and tenure) at year t, and 0 otherwise.

Table 4

Rater and Ratee Qualifications and Performance Rating

Dep. Var.: $RATING_{ijt}$		OLS
$PEER_{ijt}$	β_1	0.322***
		(0.061)
$RateeQual_{jt}$	β_2	0.188***
		(0.050)
$PEER_{ijt} \times RateeQual_{jt}$	β_3	-0.379***
		(0.103)
$RaterQual_{it}$	β_4	0.154***
		(0.040)
$PEER_{ijt} \times RaterQual_{it}$	β_5	-0.331***
		(0.098)
$RaterQual_{it} \times RateeQual_{jt}$	β_6	-0.122**
		(0.049)
$PEER_{ijt} \times RateeQual_{jt} \times RaterQual_{it}$	β_7	0.406***
		(0.135)
Rater rank fixed-effects		Yes
Ratee rank fixed-effects		Yes
Ratee fixed-effects		Yes
Year fixed-effects		Yes
Department fixed-effects		Yes
N		7,346
Adj. R^2		0.466

Notes: Robust standard errors clustered by ratee and year are reported in parentheses. *, **, *** are significant at 10%, 5%, and 1%, respectively (two-tailed).

Definition of variables:

$RATING_{ijt}$ = the rating that rater i gives to ratee j at year t, ranging from zero (poorest performance) to ten (highest performance).

$PEER_{ijt}$ = 1 if the rater i and ratee j are of the same rank at year t, and 0 otherwise.

$RateeQual_{jt}$ = 1 if the ratee j has already passed minimum requirements for promotion (in terms of attendance, academic qualifications, project experience, and tenure) at year t, and 0 otherwise.

$RaterQual_{it}$ = 1 if the rater i has already passed minimum requirements for promotion (in terms of attendance, academic qualifications, project experience, and tenure) at year t, and 0 otherwise.

Table 5

Determinants of Promotion

Dep. Var.: *PROMITION*	Logit	Logit
PR_dep	6.721***	5.472***
	(1.235)	(1.088)
RateeQual	1.546***	0.490
	(0.555)	(0.324)
PR_dep × *RateeQual*	-2.572**	
	(1.070)	
LICENSE	1.373***	1.282***
	(0.393)	(0.391)
Constant	-3.408***	-2.850***
	(0.826)	(0.784)
Rank fixed-effects	Yes	Yes
Year fixed-effects	Yes	Yes
Department fixed-effects	Yes	Yes
N	426	426
Pseudo R^2	0.380	0.370

Notes: Standard errors are reported in parentheses. *, **, *** are significant at 10%, 5%, and 1%, respectively (two-tailed).

Definition of variables:

PROMOTION = 1 if a ratee gets promoted after the annual performance evaluation scheme, and 0 otherwise.

PR_dep = the within-department percentile of an employee's average performance rating in a given year.

RateeQual = 1 if the ratee has already passed minimum requirements for promotion (in terms of attendance, academic qualifications, project experience, and tenure) in a given year, and 0 otherwise.

LICENSE = 1 if the ratee has obtained the CPA license in a given year, and 0 otherwise

Table 6

Correlation Matrix: Average Ratings from Different Components

Variables	*Rating_avg*	*Rating_head*	*Rating_peer*	*Rating_nonpeer*	*Rating_self*
Rating_avg	1.000				
Rating_head	0.834***	1			
Rating_peer	0.619***	0.520***	1		
Rating_nonpeer	0.982***	0.827***	0.490***	1	
Rating_self	0.448***	0.439***	0.448***	0.409***	1

Notes: *, **, *** are significant at 10%, 5%, and 1%, respectively.

Definition of variables:

Rating_avg = a ratee's overall average rating from department head, peers, and nonpeers.
Rating_head = a ratee's average rating from department head.
Rating_peer = a ratee's average rating from peers.
Rating_nonpeer = a ratee's average rating from nonpeers (including department head).
Rating_self = a ratee's average rating from self-evaluation.

Table 7

Summary Statistics: Counterfactual Changes of Promotion Probability

Variables	# Obs.	Mean	Median	SD	% negative	% zero	% positive
Δ_{head}	426	-0.0125	0	0.1060	38.97	28.64	32.39
Δ_{peer}	368	0.0079	0	0.1566	37.50	24.18	38.32
$\Delta_{nonpeer}$	426	0.0003	0	0.0677	27.46	47.18	25.35
Δ_{self}	402	0.0398	0	0.1398	31.84	25.37	42.79

Definition of variables:

Δ_{head} = $P_{head} - P_{actual}$.

Δ_{peer} = $P_{peer} - P_{actual}$.

$\Delta_{nonpeer}$ = $P_{nonpeer} - P_{actual}$.

Δ_{self} = $P_{self} - P_{actual}$.

P_{head} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the department head's ratings.

P_{peer} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the peer evaluation part of the 360-degree appraisal.

$P_{nonpeer}$ = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the nonpeer part of the 360-degree appraisal.

P_{self} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by self-evaluations.

P_{actual} = each employee's predicted promotion probability employing the logistic model of promotion decision in Table 5, and using the actual appraisal results.

Table 8

Correlation Matrix:

***RateeQual* dummy and Counterfactual Changes of Promotion Probability**

Variables	*RateeQual*	Δ_{head}	Δ_{peer}	$\Delta_{nonpeer}$	Δ_{self}
RateeQual	1.000				
Δ_{head}	0.017	1			
Δ_{peer}	-0.103**	0.126**	1		
$\Delta_{nonpeer}$	-0.061	0.178***	-0.283***	1	
Δ_{self}	0.033	0.304***	0.218***	0.165***	1

Notes: *, **, *** are significant at 10%, 5%, and 1%, respectively (two-tailed).

The first three rows are computed using the 368 common observations of *RateeQual*, Δ_{head}, Δ_{peer}, and $\Delta_{nonpeer}$. The last row is computed using the 345 common observations of all variables in this table. This difference is due to missing values in self-evaluation.

Definition of variables:

RateeQual = 1 if the ratee has already passed minimum requirements for promotion (in terms of attendance, academic qualifications, project experience, and tenure) in a given year, and 0 otherwise.

Δ_{head} = $P_{head} - P_{actual}$.

Δ_{peer} = $P_{peer} - P_{actual}$.

$\Delta_{nonpeer}$ = $P_{nonpeer} - P_{actual}$.

Δ_{self} = $P_{self} - P_{actual}$.

P_{head} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the department head's ratings.

P_{peer} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the peer evaluation part of the 360-degree appraisal.

$P_{nonpeer}$ = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the nonpeer part of the 360-degree appraisal.

P_{self} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by self-evaluations.

P_{actual} = each employee's predicted promotion probability employing the logistic model of promotion decision in Table 5, and using the actual appraisal results.

Table 9
Counterfactual Analysis of Promotion Outcomes

	(1) Δ_{head}	(2) Δ_{peer}	(3) $\Delta_{nonpeer}$	(4) Δ_{self}
RateeQual	0.021*	-0.032*	-0.010	0.016
	(0.012)	(0.018)	(0.008)	(0.016)
PR_dep				
LICENSE	0.012	-0.042*	0.003	-0.014
	(0.014)	(0.022)	(0.009)	(0.019)
Constant	-0.012	0.126***	0.028	0.022
	(0.027)	(0.042)	(0.018)	(0.037)
Rank fixed-effects	Yes	Yes	Yes	Yes
Year fixed-effects	Yes	Yes	Yes	Yes
Department fixed-effects	Yes	Yes	Yes	Yes
N	426	368	426	402
Adj. R^2	0.005	0.019	0.000	0.025

Notes: Standard errors are reported in parentheses. *, **, *** are significant at 10%, 5%, and 1%, respectively (two-tailed).

Definition of variables:

Δ_{head} = P_{head} - P_{actual}.

Δ_{peer} = P_{peer} - P_{actual}.

$\Delta_{nonpeer}$ = $P_{nonpeer}$ - P_{actual}.

Δ_{self} = P_{self} - P_{actual}.

P_{head} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the department head's ratings.

P_{peer} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the peer evaluation part of the 360-degree appraisal.

$P_{nonpeer}$ = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by the nonpeer part of the 360-degree appraisal.

P_{self} = each employee's counterfactual probability of promotion, which is computed employing the fitted logistic model of promotion decision in Table 5, and using the counterfactual scenario where appraisal results are determined only by self-evaluations.

P_{actual} = each employee's predicted promotion probability employing the logistic model of promotion decision in Table 5, and using the actual appraisal results.

RateeQual = 1 if the ratee has already passed minimum requirements for promotion (in terms of attendance, academic qualifications, project experience, and tenure) in a given year, and 0 otherwise.

PR_dep = the within-department percentile of an employee's average performance rating in a given year.

LICENSE = 1 if the ratee has obtained the CPA license in a given year, and 0 otherwise.