The Role of Calibration Committees in Subjective Performance Evaluation Systems

B. William Demeré
Broad College of Business
Michigan State University
270 N. Business Complex
East Lansing, MI 48824
demere@broad.msu.edu

Karen Sedatole
Broad College of Business
Michigan State University
270 N. Business Complex
East Lansing, MI 48824
sedatole@broad.msu.edu

Alexander Woods
Mason School of Business
College of William and Mary
P.O. Box 8795
Williamsburg, VA 23187
alex.woods@mason.wm.edu

May 2015

We are grateful to the representatives at our research site for their generosity in providing us data and helpful insights during interviews. We thank Jasmijn Bol, Paul Demeré, Isabella Grabner, Susan Krische, Melissa Martin, Christian Mastilak, Frank Moers, Wim Van der Stede, and Steve Wu for their very helpful comments. Lastly, this study benefited significantly from feedback from participants of the AAA Management Accounting Section Mid-Year Research Conference and workshop participants at the College of William and Mary, the Virginia Area Research Conference, the University of South Carolina, Northeastern University, Michigan State University, the University of Arkansas, American University, Southern Methodist University, and University of Kentucky.

The Role of Calibration Committees in Subjective Performance Evaluation Systems

ABSTRACT

Using proprietary data from the internal audit department of a large multinational organization, we provide the first empirical evidence of the role that "calibration committees" play in subjective performance evaluation systems. Specifically, we examine how distributional properties of ratings, supervisor rating credibility, and structural characteristics of the calibration process influence the likelihood, direction, and magnitude of calibration committee adjustments to supervisors' subjective ratings of subordinate auditors. Taken together, these results suggest that, although calibration committees do appear to make adjustments that remove inter-rater differences in the distribution of initial ratings, they have asymmetric preferences for the downward adjustment of relatively higher ratings versus the upward adjustment of relatively lower ratings. The result is that while calibration committees may mitigate leniency bias, they appear to exacerbate centrality bias. Contrary to our prediction, we find no association between calibration committee adjustments and our proxies for the credibility of supervisor ratings. However, we do find two structural characteristics of the calibration process that are associated with calibration committee adjustments. First, we find that committee adjustments are decreasing in the hierarchical distance between the committee and the subordinate auditor being rated, consistent with the collocation of decision-making authority with the individuals possessing the relevant knowledge, and, second, that adjustments are less likely when the rating supervisor serves on the committee. This study contributes to the literature on subjective performance evaluation by providing novel insights into the organizational dynamics of subjective performance evaluation systems when decision rights span hierarchical levels of the organization.

Keywords: Subjective performance evaluation; management control; calibration committee; decision rights; incentives; hierarchical distance; organizational justice.

Data Availability: A confidentiality agreement prevents the authors from distributing the data.

1. Introduction

Many organizations reward employees based on subjective performance ratings (Ittner et al. [2003], Gibbs et al. [2004], Bol [2011], Höppe and Moers [2011]). Subjectivity can improve incentive contracting by allowing supervisors to incorporate dimensions of performance not easily measured objectively and to incorporate new information into the performance evaluation process that may have unduly influenced performance in a manner beyond the control of the employee (Hopwood [1972], Baiman and Rajan [1995], Fisher et al. [2005], Rajan and Reichelstein [2009]). However, these benefits come at a cost. In particular, prior research documents negative effects associated with subjective performance evaluation, such as the tendency of supervisors to exhibit leniency and centrality biases (Murphy and Cleveland [1991], Moers [2005]), to fall prey to a halo effect (Bol [2008], Anderson et al. [2014]), or to engage in favoritism (Prendergast and Topel [1996]). A further problem arises when employees compete for shares of a common bonus pool but are subjectively evaluated by different supervisors who likely vary in their expectations; that is, on a performance scale of one to five, one supervisor's performance rating of 'five' may be another supervisor's 'three', which results in inconsistencies across supervisors (Huber [1989], Mohrman et al. [1989]).

An increasingly used management control innovation aimed at reducing the costs of subjectivity is to allocate decision rights over performance ratings to "calibration committees" typically comprised of second-level supervisors. After first-level supervisors have subjectively rated subordinate employees, these committees convene to review the subjective ratings and, where deemed necessary, make revisions (either up or down) before the ratings are disseminated back to the employees and/or used in their bonus determination (Risher [2011, 2014]).

Survey and anecdotal evidence of performance management practices indicate that it is

quite common for organizations to distribute decision rights over employees' subjective ratings across multiple levels of the organizational hierarchy, particularly in managerial and professional work settings where task complexity is high (Mohrman et al. [1989], Milkovich and Wigdor [1991]). In 2011, the Society for Human Resource Management (hereafter, SHRM) surveyed 510 organizations with 2,500 or more employees, and found that a majority (54%) of these organizations report using formal calibration committees as part of the performance evaluation process. Of those, 35 percent said they change supervisor ratings "regularly" (SHRM [2011]). Additionally, Bretz et al. [1992]—who integrated three surveys of performance evaluation practices—conclude that second-level supervisors have "significant" (p. 331) input in the final ratings for managerial and professional employees.

The primary stated purpose of calibration committees is to standardize ratings across supervisors by promoting consistent performance standards and removing or otherwise attenuating the biases that supervisors introduce.[1] While the professed benefit of calibration committees is important given the significant research documenting biases in subjective performance evaluation systems, little is known about how these committees function in practice or whether they perform the intended role of improving rating consistency.

This study provides evidence on the role of calibration committees in subjective performance evaluation systems. Specifically, our unique data allow us to provide evidence on the extent to which calibration committees achieve their oft-stated objective – namely, to provide rating consistency by removing inter-rater differences in the distribution of initial ratings. We also examine whether additional, potentially unanticipated factors related to supervisor rating credibility and structural characteristics of the calibration process are associated with the

[1] Practitioners describe numerous other benefits of the calibration process including providing a mechanism to communicate performance expectations to supervisors and facilitating more effective *ex post* performance reviews between the supervisor and the employee (e.g., Grote [2005]).

likelihood, direction, and magnitude of calibration committee adjustments.

We conduct our study using a proprietary dataset from the internal audit department of a large multinational organization collected over a three year period (2007-2009). The data consist of 1,333 initial ratings of 686 different subordinate auditors made by 110 supervisors and calibrated by 12 committees. The data provide an audit trail of the initial ratings made by supervisors and the final ratings after the calibration committees decided on adjustments. We also collect survey data from the field to provide supplemental evidence regarding supervisor and subordinate auditor perceptions of justice associated with the calibration process.

Our first objective is to provide evidence about the extent to which calibration committees make adjustments to improve the consistency of ratings across different supervisors. We find that calibration committees are about three times more likely to adjust extreme high initial ratings that would result in unusually high bonus payouts, which may reflect favoritism on the part of the supervisor toward an employee. However, we do not find an increased likelihood of adjustment for extreme low initial ratings. We also find that the committees are about three times more likely to adjust ratings given by supervisors with relatively high mean ratings as compared to other supervisors, but not more likely to adjust ratings made by supervisors with relatively low mean ratings. Taken together, these results suggest that, although calibration committees do appear to make adjustments that remove inter-rater differences in the distribution of initial ratings, they have asymmetric preferences for the downward adjustment of relatively high ratings versus the upward adjustment of relatively low ratings. The result is an adjusted (i.e., calibrated) rating distribution that has a lower mean, possibly reflecting the committees' efforts to reduce the leniency bias. However, the adjusted rating distribution also has a lower standard deviation relative to the distribution of initial ratings. Thus, while calibration

committees appear to mitigate leniency bias, they exacerbate centrality bias.

Prior research suggests decision-makers will place greater reliance on representations deemed to be more credible (Pornpitakpan [2004], Dholakia and Sternthal [1977], Birnbaum and Stenger [1979]). Our second objective is to thus examine whether proxies for the credibility of a supervisor rating is negatively associated with the calibration committee's likelihood of adjusting that rating. Rating credibility proxies include supervisor own performance, supervisor span of control (i.e., the number of subordinates being supervised, suggesting the opportunity to benchmark subordinates against one another), and the standard deviation of the supervisor's ratings (indicating the supervisor's ratings are, indeed, reflective of performance differences). We find no relation between the likelihood that the calibration committees will adjust supervisor ratings and any of our three proxies for rating credibility.

Finally, we examine two structural characteristics of the calibration process: (i) the hierarchical distance (the number of layers of authority) between the committee and the subordinate being evaluated, and (ii) whether the rating supervisor serves on the committee. We find that as hierarchical distance increases, the likelihood that the committee will adjust the supervisor rating decreases, and the magnitude of any such adjustment is lower. This finding supports the theoretical prediction that decision rights are optimally collocated with the knowledge relevant for making a particular decision; that is, when the supervisor has a large information advantage relative to the calibration committee (i.e., when there is greater hierarchical distance between the committee and the subordinate), the committee is more likely to defer rating judgment to the supervisor by not adjusting the rating. We also find that calibration committees are about 0.7 times as likely to adjust the ratings of a supervisor serving on the calibration committee.

This study contributes to the literature by providing novel insights into the organizational dynamics of subjective performance evaluation systems when performance evaluation decision rights span hierarchical levels of the organization. More specifically, we make two important contributions. First, to our knowledge, the study provides the first evidence of the role of calibration committees in subjective performance evaluation systems. In doing so, we document both intended effects such as improved cross-sectional rating consistency, as well as unintended effects in the form of exacerbated centrality bias and the hesitancy of committee members to adjust the ratings of supervisors that also serve on the committee. Despite the recognized importance of gaining a better understanding of this practice (Harris [1994], Arvey and Murphy [1998], Prendergast and Topel [1993]), there is virtually no empirical evidence on this subject, primarily due to the unavailability of data (Prendergast [1999]).

Second, this study sheds light on the reality that immediate supervisors do not always have exclusive decision rights over the ratings of subordinate employees. Prior archival research on subjective rating biases (e.g., leniency and centrality) examines final, potentially adjusted, subjective ratings. Our results show that, because calibration committees appear to reduce leniency bias, the bias of immediate supervisors may be higher than previous research suggests. Paradoxically, we also show that input from higher level sources may be more responsible than immediate supervisors for increasing centrality bas. These insights are gleaned only by separately examining the initial ratings made by supervisors and the final, adjusted ratings.

The remainder of the paper is organized as follows: Section 2 formalizes the research hypotheses, Section 3 describes the research setting, Section 4 describes the variable measurements, Section 5 presents the empirical models and results, and Section 6 concludes.

2. Hypothesis Development

An organization's control system provides a framework for planning, controlling, and monitoring activities (Chenhall [2003], Coletti et al. [2005]) and includes the assignment of decision-making authority (Moers [2006], Jensen and Meckling [1992]). Because supervisors usually possess significant information about their subordinates' performance not otherwise captured by available quantitative performance measures, organizations frequently authorize supervisors to subjectively rate their subordinates' performance (Ittner et al. [2003], Gibbs et al. [2004], Bol [2008, 2011], Höppe and Moers [2011]). This allows supervisors to incorporate dimensions of performance not easily measured objectively and to incorporate new information into the performance evaluation process that may have unduly influenced performance in a manner beyond the control of the employee (Hopwood [1972], Baiman and Rajan [1995], Fisher et al. [2005], Rajan and Reichelstein [2009]). This can result in improved contract efficiency by reducing risk imposed on the employee and, hence, the risk premium that must be paid.

Although subjective evaluation has many benefits, subjective evaluation systems suffer from at least three limitations. First, supervisors often fall prey to biases in making their subjective ratings; namely, supervisors tend to exhibit both a leniency and a centrality bias (Saal and Landy [1977], Murphy and Cleveland [1991]). Leniency bias refers to the pervasive and well-documented tendency of supervisors to rate an employee higher than what the employee's performance warrants, and arises from uncertainty in performance evaluation coupled with asymmetrical preferences of supervisors to inflate ratings rather than to potentially understate ratings (Bol [2011], Golman and Bhatia [2012]). Centrality bias refers to the tendency for ratings to be compressed around some value, resulting in a narrow distribution of ratings and a lack of differentiation between ratings (Murphy and Cleveland [1991], Moers [2005]).

Second, different supervisors are likely to have different performance standards, resulting in inconsistencies across ratings of different supervisors (Huber [1989], Mohrman et al. [1989]). That is, two employees with exactly the same level of performance but working for different supervisors may be given different ratings. This is especially problematic when these employees compete for shares of a common bonus pool. Third, supervisors may intentionally apply standards inconsistently to their own subordinates, using ratings to favor some subordinates over others, or to discriminate against some subordinates (Prendergast and Topel [1996]).

To overcome these inherent limitations in subjective performance evaluation, while still promoting the benefits, organizations often allocate decision rights over subjective performance ratings across multiple levels of the organizational hierarchy through the use of "calibration committees" (Risher [2011, 2014], Merchant [1989], Mohrman et al. [1989], Milkovich and Wigdor [1991]). These committees review supervisors' subjective ratings of subordinates and, when deemed necessary, adjust (either up or down) the initial ratings that supervisors assign (e.g., Grote [2005], McGregor [2006]). Typically the stated objective of such committees is to achieve cross-rater consistency in the assignment of ratings, thereby promoting the effectiveness, fairness, and integrity of the performance evaluation and incentive system (Merchant [1989], Bretz et al. [1992], Grote [2005], McGregor [2006]).

Below we develop hypotheses regarding three categories of determinants of calibration committee adjustments: (1) distributional properties of the initial ratings, (2) factors suggesting supervisor rating credibility, and (3) structural characteristics of the calibration process.

2.1 DISTRIBUTIONAL PROPERTIES OF INITIAL RATINGS

Because one of primary purposes of the calibration committee is to address supervisor biases and improve cross-rater consistency in the assignment of ratings, we first theorize that

committees will base adjustment decisions, at least in part, on distributional properties of initial ratings. By nature of their access to all ratings for a subset of employees, the calibration committee has superior statistical information about the overall distribution of ratings and differences in mean ratings across supervisors. Consistent with the stated objective to mitigate bias and improve cross-rater consistency, adjustments made based on observed distributional properties can promote actual and perceived "organizational justice" or "fairness" in the subjective performance evaluation system (Schappe [1998], McFarlin and Sweeney [1992], Colquitt [2001], Colquitt et al. [2001], Colquitt et al. [2013]).

We first examine ratings that lie at the extremes of the subjective rating scale, either low or high. While an extreme high (low) rating might be an accurate assessment of a highly (poorly) performing subordinate, it is also possible that extreme ratings are the result of a supervisor's intentional effort to single out individual subordinates for purposes of *unfairly* favoring them or discriminating against them (Bjerke et al. [1987], Prendergast and Topel [1996], MacLeod [2003]). Even if an extreme rating is not reflective of intentional supervisor bias (either favoritism or discrimination), such a rating is likely to be perceived as such by others within the organization. We thus posit that calibration committees will be more likely to adjust extreme high ratings and extreme low ratings relative to moderate ratings order to reduce (real or perceived) favoritism and discrimination, and promote perceptions of organizational justice. Specifically, we make the following prediction:

H1a: Calibration committees are more likely to adjust extreme *high* and extreme *low* supervisor ratings, relative to moderate ratings.

While hypothesis H1a predicts calibration committees adjust extreme individual ratings of specific employees, hypothesis H1b relates to the possibility that supervisors will vary in their degree of leniency (or stringency), *on average*. According to the SHRM [2011] survey, the most

frequently cited reason for adjusting ratings (69%) is to "fix" those ratings deemed to be inconsistent across different supervisors. Grote [2005] also notes that "a growing number of companies" (p. 153) are using a review, or "calibration", process to ensure that employees are evaluated similarly, even though they work for different supervisors. Thus, a primary purpose of reviewing and calibrating supervisors' subjective ratings is to ensure that employees working for different supervisors are evaluated similarly (e.g., Grote [2005], McGregor [2006]); that is, to mitigate *differences* in leniency (or stringency) across supervisors that undermine perceptions of organizational justice.

Because subjective performance evaluation is inherently noisy, especially in high task complexity settings, supervisors must weigh the possibility of committing favorable errors (i.e., ratings that are undeservedly high) against unfavorable errors (i.e., ratings that are undeservedly low). Prior research suggests that supervisors prefer favorable over unfavorable errors. Their aversion toward unfavorable errors can reflect sympathy for, or loyalty to, an employee (Giebe and Gürtler [2012]), and reflects a desire to avoid conflict with a disgruntled employee (Prendergast and Topel [1996], Prendergast [2002]). Even if the supervisor desires to accurately rate the employee, because of the inherent uncertainty in the evaluation, the supervisor's preference for favorable errors will lead to ratings that are, on average, too high (i.e., lenient) (Golman and Bhatia [2012]).

We consider the possibility that there is variation across supervisors in the extent of asymmetrical preferences for favorable errors. Because the leniency bias increases in the degree of asymmetrical preferences for favorable errors (Bol [2011], Murphy and Cleveland [1991]), supervisors with stronger asymmetrical preferences will tend to be more lenient and give higher mean ratings relative to other supervisors, while those with weaker asymmetrical preferences

will be relatively more stringent and tend to give lower mean ratings (which may still be inflated relative to true performance) (Saal et al. [1980]). Kane et al. [1995] provide field evidence across three different field settings that the degree of leniency is a stable individual tendency. Consistent with this, Vance et al. [1983] investigates supervisor-specific rating variation over five successive rating periods using a methodology that allows them to estimate how much rating variance is attributable to supervisor as compared to employee characteristics (e.g., ability). They conclude that "the majority (about four times as much) [variance in supervisory mean ratings] may be attributed to raters" (p. 617).

We posit that calibration committees will infer variation in supervisor leniency from supervisor mean ratings and will use adjustments to remove this source of cross-rater variation in subjective ratings. More specifically, committees are more likely to adjust the ratings of supervisors deemed to be more lenient (stringent) as indicated by a high (low) mean rating relative to the rating scale midpoint. Stated formally:

H1b: Calibration committees are more likely to adjust ratings made by supervisors giving *high (low)* mean subjective ratings, relative to supervisors giving mean ratings closer to the rating scale midpoint.

Finally, we make a summary prediction that calibration committees, recognizing the general propensity for leniency (as opposed to stringency) in supervisor subjective ratings overall, will be more likely to downward adjust supervisor ratings than to upward adjust those ratings, on average. Further, the magnitude of downward adjustments will be larger than the magnitude of upward adjustments, on average. Stated formally:

H1c: Calibration committees are more likely to downward adjust supervisor ratings than to upward adjust those ratings, on average, and the magnitude of downward adjustments will be *larger* than the magnitude of upward adjustments, on average.

2.2 SUPERVISOR RATING CREDIBILITY

We also consider whether calibration committees will vary adjustment decisions depending on cues regarding the credibility of the initial ratings. We consider three such credibility cues. First, based on prior research (e.g., Pornpitakpan [2004], Dholakia and Sternthal [1977], Beaulieu [1994, 2001], Birnbaum and Stenger [1979], Morhman et al. [1989], Spence and Keeping [2010]), we expect supervisor own performance to increase perceptions of supervisor credibility, which will then inspire greater confidence in, and reliance upon, initial subjective ratings. In particular, prior research documents that more experienced and more skilled supervisors provide subordinate ratings that are more accurate (Schneider and Bayroff [1953], Morhman et al. [1989]), more objective (Tan and Jamal [2001]), and less lenient (Spence and Keeping [2010]). Second, we expect calibration committees to view ratings to be more credible when made by supervisors who are tasked with evaluating multiple subordinates because such supervisors have greater access to performance benchmarking information (i.e., they can compare performance across subordinates). Lastly, we expect calibration committees will assess ratings from supervisors who make greater distinctions across subordinates to be more credible. In sum, we hypothesize calibration committees will use cues to assess the credibility of initial ratings, and will be less likely to adjust ratings that those cues suggest are more credible. Stated formally:

H2: Calibration committees are *less* likely to adjust supervisor ratings in the presence of cues that suggest rating credibility.

2.3 STRUCTURAL CHARACTERISTICS OF THE CALIBRATION PROCESS

Finally, we examine two structural characteristics of the calibration process: the hierarchical distance (i.e., the number of layers of authority, Liberti and Mian [2009]) between the calibration committee members and the subordinates being evaluated, and whether a

subordinate's supervisor serves as a member of the calibration committee. Optimal decision-making requires the collocation of decision-making authority with the individual possessing the knowledge relevant for making a particular decision (Jensen and Meckling [1992]). As hierarchical distance between the calibration committee members and the subordinate being rated increases, committee members tend to possess less specific knowledge about subordinate performance, and the rating supervisor has a greater information advantage relative to the committee regarding subordinate performance. In such circumstances, we expect the calibration committee to defer decision-making authority for rating subordinate performance to the supervisor. That is, we predict that calibration committees will be less likely to adjust supervisor ratings for subordinates further removed from the committee in the organizational hierarch. This leads to the following hypothesis:

H3a: Calibration committees are *less* likely to adjust supervisor ratings when the hierarchical distance between the subordinate and the calibration committee is larger.

Importantly, hypothesis H3a is in contrast to hypotheses H1a-c which predict that the calibration committee has superior statistical knowledge regarding the *distribution* of initial ratings, and thus will be more likely to make adjustments related to the distribution of ratings. Indeed, Grant [1996] emphasizes that "the principle of collocation requires that decisions based upon [specific] knowledge are decentralized, while decisions requiring statistical knowledge are centralized" (p. 119).

We also consider whether having a supervisor serve as a member of a calibration committee will affect the likelihood of the committee adjusting ratings given by that supervisor. We posit that the calibration committee will be less likely to adjust the ratings of supervisors who are also members of the committee. First, supervisors who are also committee members will devote more time and attention in assigning their initial ratings so as to increase their ability to

justify those ratings before their fellow committee members (Cardinaels and Labro [2008]). That is, the knowledge that face-to-face discussions with the calibration committee will ensue could provide a disciplining mechanism for the supervisor to make more justifiable ratings to begin with. Second, face-to-face discussions with committee members provide increased opportunity for lobbying on the part of the supervisor. Fellow committee members may be less willing to question the rating in such face-to-face exchanges, thereby avoiding direct conflict. This leads to our final hypothesis:

H3b: Calibration committees are *less* likely to adjust ratings given by supervisors who also serve as members of the calibration committee.

3. Research Setting

The research setting for this study is the internal audit department of a large multinational organization with five hierarchical levels of auditors: junior auditors, staff auditors, senior auditors, supervisory auditors, and program directors. To gain insight into the organizational setting and to facilitate interpretation of our results, prior to gathering data and conducting analyses we interviewed and surveyed personnel spanning all levels of the organizational hierarchy, from both headquarters and from "the field." We also obtained and reviewed various documents about the organization, the performance evaluation and rewards system, and the subjective rating calibration process.

Generally, supervisory auditors rate the junior, staff, and senior auditors, and program directors rate the supervisory auditors.[2] Junior, staff, and senior auditors are generally geographically disbursed across the organization's 22 worldwide field offices, while program directors and senior executives are generally located at either the organization's headquarters or

[2] There are a few exceptions. A program director—not a supervisory auditor—was the supervisor for three staff auditors and six senior auditors. A senior executive—not a program director—was the supervisor for 19 supervisory auditors.

operations center. The organization insists that the nature of internal audit work renders them unable to adequately anticipate and pre-define the cost, quality, quantity, results, or timeliness of the work that needs to be accomplished. The increasing complexity and nature of internal audit work has only exacerbated this difficulty (McDonald [2003]). As a result, junior, staff, and senior auditors are rated on four subjective measures, while supervisory auditors and program directors are rated on five subjective measures. All ratings are on a scale of 1 to 5, and an overall subjective rating is computed from a weighted average of the individual ratings. The overall subjective ratings are calibrated by the calibration committees (described below).

The organization uses bonus pools to distribute performance pay to auditors based on the calibrated overall subjective ratings. The bonus pools operate similarly to how prior literature has described bonus pools (c.f., Rajan and Reichelstein [2006, 2009]). Specifically, the organization pre-commits to a fixed amount of funds for a separate bonus pool at each of the five levels of the organizational hierarchy, and funds are not transferred across pools. Each "share" of a given pool represents a percentage of that pool, and the value of each share is inversely related to the total number of shares within the pool.[3] Auditors receive between 0 and 6 shares in their respective bonus pools, according to a pre-defined sharing function with seven subjective rating performance levels (Figure 1, Panel A). An individual auditor's performance pay is a function of the number of shares he receives, the value of each share (a function of the total number of shares in the pool), and his salary. The details of the pay plan are transparent to all employees.

3.1 CALIBRATION PROCESS OVERVIEW

The organization believes that inconsistencies in supervisors' ratings jeopardize the fairness and incentive effects of their evaluation system. Thus, as part of its performance

[3] Each pool's share value = [Total pool funding/($\sum_{i=1}^{n} (x_i \times y_i)$], where x_i and y_i are the salary and assigned shares, respectively, of each employee in the pool. Untabulated analyses indicate that there are fewer total shares following the calibration process (after adjustments) for every pool in every year.

evaluation system, the organization uses calibration committees in an attempt to ensure that all employees are evaluated similarly, even though they work for different supervisors. While it is true that committee members generally do not observe supervisors' private information about the individual auditors whose performance ratings they are reviewing,[4] it is also true that supervisors do not observe how the performance ratings of auditors they evaluated compare to auditor ratings provided by other supervisors. The organization believes the calibration committee has this macro-level knowledge. That is, the committees assess the performance ratings of auditors *across* supervisors. By identifying when differences in subjective ratings have occurred, the organization believes the calibration process increases the fairness of the performance ratings, which promotes the integrity and incentive effects of the entire subjective pay-for-performance measurement system. Indeed, organizational documents state that the purpose of the committee is to "…ensure that the same standards for evaluating performance are applied."[5]

There is a separate calibration committee for each level of the hierarchy for each year; each committee is composed of three to five committee members. In our data, we have a total of 12 calibration committees (five committees each for years 2007 and 2009, two committees for 2008).[6] The composition of the calibration committee for each hierarchical level can change from one year to the next (range of 0% – 60% in committee member composition change), and no committee member serves on more than one committee in a given year. Calibration committee members are not directly compensated for their review of subjective performance

[4] Infrequent exceptions occur when one of the calibration committee members is in an employee's direct line of authority (as the employee's first- or second-level supervisor).

[5] Organizational documents state that "The process provides a structured way of reviewing supervisor decisions, preserves the integrity of the performance management system, and ensures the fairness of the system." Documents explicitly define "fair" and intertwine the definition with maximizing the incentive effect of the pay-for-performance system. Specifically, documents say that "The performance management system aims to compensate the workforce fairly by appropriately rewarding employees for the contributions they make" and that *actual* fairness is achieved when "People who contribute more to (the organization) get more".

[6] We were unable to collect performance data for staff and senior auditors for 2008.

ratings, nor are they participants in the same bonus pools of the auditors whose ratings they are potentially adjusting. Serving as a committee member is an "additional duty" for which committee members receive no explicit rewards. Nevertheless, they do have implicit incentives to perform their duty effectively because they occupy positions of authority with resource oversight and promotion responsibility for the subordinate auditors. Moreover, because committee members typically occupy positions higher than the supervisors in the organization hierarchy, they are at least one level removed from some of the incentives that first-level supervisors might have to bias ratings.[7]

3.2 CALIBRATION PROCESS DETAILS

After supervisors have prepared their initial "recommended" ratings, each calibration committee meets in what the organization calls "review meetings." During the meetings, the committee considers the supervisors' written summaries of the accomplishments and performance of his subordinate auditors. They deliberate until they achieve a "common understanding" of the types of auditor achievements and contributions that warrant various performance ratings for the auditors being considered by that committee. Once a common understanding is reached, the committee members use this to guide their adjustments to supervisors' initial ratings. Each individual adjustment must be made with unanimous agreement; if consensus cannot be achieved (a rare event), then the determination is made at the next highest level in the organization. If they choose, the committee may discuss a performance rating with the rating supervisor and request additional information. These elements of the calibration process are designed to provide the calibration committee with a fine-tuned understanding of how the auditors were evaluated.

The calibration process culminates in the calibration committee's decision to either adjust

[7] There are some exceptions when the subordinate auditor supervisor also serves on the committee.

(upward or downward) or not adjust the initial ratings made by supervisors. Ultimately, decision-making authority over final ratings resides with the committee. After the committees' adjustments are complete, supervisors hold meetings with auditors to discuss their ratings. Although the process itself is publicized and transparent, auditors are not privy to either their supervisor's initial rating or the details that produced their final rating; only their final rating is shared (supervisors are not permitted to share ratings with auditors until the calibration committee has approved them).

Aside from the above controls, there are no pre-determined quotas or forced distributions that the calibration committee must adhere to.[8] However, there is an implicit expectation that "most" auditors who make valuable contributions to the organization will receive ratings of 3-4 on the one to five scale (earning between 2 and 3 bonus shares, see Figure 1, Panel A), and that only a "select few" will truly be performing at the highest performance levels with ratings above 4 (which earn between 4 and 6 bonus shares). The organization opted to use a calibration system over a forced distribution system because it believes the latter would undermine the goal of fairly assessing auditor performance.

4. Data Description and Variable Measurements

We gathered data from our research site on the supervisor initial ratings, the final adjusted (i.e., calibrated) ratings, auditor salary, and bonus payouts. Data are from 2007-2008 and consist of 1,333 initial ratings of 686 different subordinate auditors made by 110 supervisors and calibrated by 12 committees.[9] Figure 1, Panel B, shows how calibration committees,

[8] Organizational documents state (and untabulated analysis confirms) that "Pre-established limits for the percentage or number of ratings that may be assigned are not permitted."

[9] These data are part of a larger dataset used in the following other studies: Sedatole et al. (2015) examine the effects of horizontal monitoring and team member dependence on performance, Demeré et al. (2015) examine the incentive effects of relative performance evaluation as a function of the *ex ante* probability of promotion, and Reichert and Woods (2015) examine behavioral motivation orientations. Our study examines calibration committees, which are

supervisors, and subordinate auditors fit within the organization's hierarchy.

4.1 DEPENDENT VARIABLES

Our primary dependent variable of interest is an indicator variable set equal to one for supervisor subjective ratings that are adjusted (either up or down) by the calibration committee, and zero otherwise (*Adjustment*). We also provide evidence regarding two additional dependent variables of interest. The first is a multilevel categorical indicator variable that represents the direction of adjustments (DOWN, NONE, and UP defined as -1, 0, and 1, respectively) (*Adj_Direction*), and the second is the absolute magnitude of the committee adjustment (*Adj_Abs_Magnitude*).

4.2 INDEPENDENT VARIABLES

H1a predicts calibration committees will be more likely to adjust extreme high and extreme low ratings. Extreme high ratings (*Extreme_High)* are coded 1 if the supervisor's initial rating is above 4 (in the 4-6 bonus share range), and 0 otherwise; extremely low ratings (*Extreme_Low*) are coded 1 if the supervisor's initial rating is below 3 (in the 0-1 share range), and 0 otherwise.[10] H1b predicts calibration committees will be more likely to adjust ratings of supervisors whose mean ratings differ from the rating scale midpoint (i.e., 3 on the 1-5 scale). *High_Sup_Avg* is coded 1 for all ratings made by supervisors with a mean initial rating above 4.00, and 0 otherwise. *Low_Sup_Avg* is coded 1 for all ratings made by supervisors with a mean initial rating below 3.00, and 0 otherwise.[11]

H2 predicts calibration committees will be less likely to adjust ratings in the presence of

---

not considered in any of these other studies. The data used in the current paper comes from a different data set than that of Woods (2012).

[10] A total of 312 ratings are classified as extreme, with 246 (18 percent of the total) in the 4-or-more share range, and 66 (5 percent of the total) in the 0-1 share range.

[11] A total of 202 ratings are classified as being assigned by a supervisor either a high or low mean rating, with 156 (12 percent of the total) coming from a supervisor with a high mean rating, and 46 (3 percent of the total) coming from a supervisor with a low mean rating.

cues that suggest rating credibility. We use three proxies for rating credibility: (i) the supervisor's own performance (*Sup_Perf*), (ii) the supervisor's span of control, measured as the number of subordinate auditors evaluated by that supervisor in a particular year (*Sup_Span*), and (iii) the standard deviation of supervisor subordinate ratings within a year (*Sup_SD*).

H3a predicts that calibration committees will be less likely to adjust supervisor ratings of subordinate auditors for whom the hierarchical distance between the subordinate and the committee is greater. Accordingly, we use as our measure of hierarchical distance, *Hier_Dist*, the number of levels of authority between the subordinate auditor and the calibration committee (See Figure 1, Panel B). For example, the calibration committees for ratings of junior auditors are composed of supervisory auditors who are three levels above the junior auditors (i.e., *Hier_Dist* = 3). In contrast, the committees for ratings of program directors are composed of senior executives who are only one level higher in the organizational hierarchy (i.e., *Hier_Dist* = 1).

Finally, H3b predicts that calibration committees will be less likely to adjust ratings of a supervisor who serves on their calibration committee. Thus, we include an indicator variable (*Sup_Member*) that takes the value of 1 if one of the calibration committee members is in the auditor's direct line of authority (as either the first-level supervisor or second-level supervisor), and 0 otherwise.

5. Models and Results

5.1 DESCRIPTIVE STATISTICS

Before presenting the results of our hypothesis tests, Tables 1 and 2 provide various descriptive statistics. Table 1, Panel A shows the mean and standard deviations of initial and adjusted ratings for each of the 12 calibration committees and in total. Table 1, Panel B, provides a frequency distribution of initial ratings in total and by calibration committee and reveals that

1,021 (77%) of the 1,333 supervisor initial ratings lie between 3.00 and 4.00 (i.e., the 2 and 3 bonus share categories). Table 1, Panel C provides a frequency distribution of calibration committee adjustments in total and by committee and shows that 327 (25%) ratings were adjusted. The committees lowered supervisor initial ratings (262 times, or20%) more than four times more often than they raised them (65 times, or 5%). Figure 2 provides a graphical representation of the rating adjustments (y-axis) plotted against the initial ratings (x-axis). The figure again shows the asymmetry in calibration adjustments, with higher initial ratings being adjusted more frequently than lower initial ratings. Table 2 provides descriptive statistics of the variables in the main analysis (Panel A) and reports correlations (Panel B).

5.2 SPECIFICATION OF MODELS

Prior to testing our hypotheses regarding calibration committee adjustments, for completeness and to facilitate interpretation of our tests of hypotheses, we first examine the determinants of supervisor initial subjective ratings. We estimate the following model using OLS regression for auditor rating *i*, supervisor *j*, and year *t*:

$$\begin{aligned} Initial\_Rating_{ijt} = {} & \beta_0 + \beta_1 Sup\_Perf_{jt} + \beta_2 Sup\_Span_{jt} + \beta_3 Sup\_SD_{jt} + \beta_4 Hier\_Dist_{it} \\ & + \beta_5 Sup\_Member_{ijt} + \beta_6 Certification_{it} + \beta_7 Ineligible_{it} + \beta_8 Hier\_Level_{it} \\ & + [Year\ indicators] + e_{ijt} \end{aligned} \quad (1)$$

In addition to the variables hypothesized to affect calibration committee adjustments—*Sup_Perf, Sup_Span, Sup_SD, Hier_Dist, and Sup_Member* (excepting the extreme high/low and mean supervisor high/low variables)—we include year indicators and three auditor-level determinants of performance. First, we include *Certification* which is an indicator variable that equals 1 if an auditor has a professional certification (e.g., CPA, CIA), and 0 otherwise. Second, based on prior literature that examines the important role that promotions can play in motivating performance (Demeré et al. [2015], Campbell [2008]), we include *Ineligible* which is an

indicator variable that equals 1 if an auditor is ineligible for promotion in a given year, and 0 otherwise. Lastly, we include *Hier_Level* which is the subordinate auditor's level in the organizational hierarchy, where 1 = junior auditor and 5 = program director (see Figure 1).

Equation (2) provides the model used to test our hypotheses:

$$[DV]_{ijt} = \beta_0 + \beta_1 Extreme\_High_{ijt} + \beta_2 Extreme\_Low_{ijt} + \beta_3 High\_Sup\_Avg_{jt} + \beta_4 Low\_Sup\_Avg_{jt} + \beta_5 Sup\_Perf_{jt} + \beta_6 Sup\_Span_{jt} + \beta_7 Sup\_SD_{jt} + \beta_8 Hier\_Dist_{it} + \beta_9 Sup\_Member_{ijt} + [Calibration\ committee\ indicators] + e_{ijt} \quad (2)$$

where subscripts *ijt* indicate auditor rating *i*, supervisor *j*, and year *t*. We estimate equation (2) for our primary dependent variable of interest, *Adjustment*, using logistic regression. Hypothesis H1a predicts positive coefficients (odds ratios > 1) for β_1 and β_2. Hypothesis H1b predicts positive coefficients (odds ratios > 1) for β_3 and β_4. Hypothesis H2 predicts negative coefficients (odds ratios < 1) for β_5, β_6, and β_7. Hypotheses H3a and H3b predict negative coefficients (odds ratios < 1) for β_8 and β_9, respectively.

In addition, to gain insights regarding the direction of calibration committee adjustments, we use multinomial logistic regression to estimate equation (2) using *Adj_Direction* as the dependent variable. Finally, to examine the effects of our predictors on the absolute magnitude of adjustments, we use *Adj_Abs_Adjustment* as the dependent variable and estimate equation (2) using OLS regression. For all models, p-values are computed from standard errors clustered by calibration committee.[12] As a reference, Table 3 provides descriptive statistics for each type of adjustment ("DOWN", "NONE", or "UP").

[12] Several options exist to analyze data that have dependencies in observations (Petersen [2009], Gow et al. [2010]). In reported analyses, we use Huber-White cluster-corrected standard errors, clustering on calibration committee because the unit of analysis is the committee members' choice of whether to adjust or not to adjust a supervisor's rating of an auditor. In addition, many employees are evaluated by different supervisors or move to another bonus pool across years, and the composition of the calibration committee changes across years. For these reasons, error terms are not likely correlated for auditors. Nevertheless, inferences are robust to additional methods of clustering, including clustering by supervisor and auditor.

5.3 RESULTS OF MODEL ESTIMATIONS

The results of estimating equations (1) and (2) are presented in Table 4. Column 1 provides the results of estimating equation (1) (i.e., the determinants of initial ratings). We do not make specific predictions regarding the determinants of initial ratings, but we find that higher auditor initial ratings are given by higher performing supervisors (0.176, $p < 0.01$) and supervisors that make greater distinctions among subordinate auditors as evidenced by larger standard deviations (0.313, $p < 0.01$). Initial ratings are lower for subordinate auditors who have greater hierarchical distance from the calibration committee (-0.144, $p < 0.10$) and lower for auditors that are ineligible for promotion (-0.200, $p < 0.01$).

Column 2 presents the results of estimating equation (2) using *Adjustment* as the dependent variable and provides the tests of hypotheses H1a, H1b, H2, H3a, and H3b. Column 3 presents the multinomial logistic estimation of equation (2) using *Adj_Direction* as the dependent variable and reports a comparison of upward adjustments relative to the base case of a downward adjustment. Column 4 presents the results of estimating equation (2) using OLS and with the absolute magnitude of adjustments, *Adj_Abs_Magnitude,* as the dependent variable. Columns 5-7 repeat the estimations reported in columns 2-4 but with the inclusion of *Initial_Rating* as an additional control variable.[13] Regression coefficients are reported for OLS estimations (columns 1, 4, and 7), and odds ratios are reported for the logistic and multinomial logistic models (columns 2-3, and 5-6).

5.3.1 Hypotheses H1a, H1b, and H1c

H1a predicts that calibration committees are more likely to adjust extreme initial ratings (*Extreme_High* = 1 and *Extreme_Low* = 1). We find that calibration committees are over three

[13] To avoid multicollinearity and to ease interpretation, *Extreme_High* and *Extreme_Low* are omitted from the models that include *Initial_Rating*.

times more likely to adjust extreme high initial ratings as compared to moderate ratings (column 2, odds ratio of 3.051, $p < 0.01$). Not surprisingly, we find that calibration committees are 0.2 times as likely to upward adjust extreme high initial ratings as they are to downward adjust those ratings (column 3, odds ratio of 0.219, $p < 0.01$). Lastly, the absolute magnitude of such committee adjustments is larger than for non-extreme initial ratings (column 4, 0.071, $p < 0.01$). Thus, to the extent that favoritism is present in initial ratings, the calibration committees appear to attenuate this problem. However, we find that calibration committees are no more likely to adjust extreme low ratings as they are to adjust non-extreme ratings (column 1, odds ratio of 0.711, $p > 0.10$). We also find no difference in the propensity to upward adjust versus downward adjust those ratings (column 3), nor in the adjustment magnitudes (column 4). Based on these results, it appears that calibration committees have an asymmetric response to extreme initial ratings, where ratings that are extreme high are more likely to be adjusted, on average, but extreme low ratings are no more likely to be adjusted than other ratings.

H1b predicts calibration committees will be more likely to adjust ratings of supervisors whose mean ratings are higher than the rating scale midpoint (*High_Sup_Avg* = 1) or lower than the midpoint (*Low_Sup_Avg* = 1). Our results show calibration committees are over three times more likely to adjust ratings of supervisors with high mean ratings (column 2, odds ratio of 3.071, $p < 0.01$), marginally less likely to adjust those ratings upward versus downward (column 3, odds ratio of 0.447, $p = 0.11$), and that the absolute magnitude of such adjustments is larger, on average (column 4, 0.070, $p < 0.01$). We find no difference in the calibration committees' likelihood of adjusting initial ratings or in adjustment magnitudes for supervisors with low mean ratings. We do find that upward adjustment is much more likely than downward adjustment for supervisors with low mean ratings; indeed, zero downward adjustments were made to ratings of

these supervisors. As with extreme individual ratings, calibration committees appear to exhibit a similar asymmetric response to mean ratings across supervisors, where ratings of supervisors with a high mean rating are more likely to be adjusted, on average, but ratings of supervisors with a low mean rating are no more likely to be adjusted than those of supervisors with mean more closer to the rating scale midpoint.

Hypothesis H1c makes a summary prediction that downward adjustments of supervisor ratings will be more common than upward adjustments, on average, and the magnitude of downward adjustments will be larger than the magnitude of upward adjustments, on average. Consistent with this hypothesis and with results reported in Table1, Panel C, we find that calibration committees make more downward adjustments (262) than upward adjustments (65), and that the difference in proportion is statistically significant ($z = 11.63$, $p < 0.001$) (untabulated). Further, we find a larger absolute magnitude for downward adjustments (0.310) relative to upward adjustments (0.283) (untabulated), however the difference is not statistically significant ($t = 1.24$, $p = 0.11$, one-tailed).

We also examine the effect on the final rating distribution that results after calibration committee adjustments (see Table 1, Panel A). We find that the sample mean final (adjusted) rating of 3.511 is significantly lower than the mean initial rating of 3.559 ($t = 10.541$, $p < 0.01$). Indeed, in 9 out of the 12 committees, the mean adjusted rating is significantly lower than the initial rating (Table 1, Panel A). We also find that the standard deviation of the final (adjusted) ratings (0.552) is significantly lower than the standard deviation of the initial ratings (0.590) ($F = 1.14$, $p < 0.05$).[14] Collectively, this evidence provides strong support for H1c. Further, as assertions of leniency and centrality bias are often made on the basis of observing rating

[14] In untabulated analyses, we also find that there are fewer total shares following the calibration process (after adjustments) for every bonus pool in every year.

distributions with a high mean and low standard deviation, respectively (Murphy & Cleveland [1991]), we document that, to the extent the original distribution of initial (pre-adjusted) ratings exhibited leniency and centrality biases, the calibration committee reduces leniency bias but exacerbates centrality bias.

5.3.2 Hypothesis H2

H2 predicts calibration committees will be less likely to adjust ratings that are more credible, where rating credibility is proxied by *Sup_Perf*, *Sup_Span*, and *Sup_SD*. Contrary to H2, we find that calibration committees are no less likely to adjust ratings for supervisors with higher performance (column 2, *Sup_Perf* odds ratio of 0.864, $p > 0.10$), with larger span of control (*Sup_Span*, odds ratio of 0.984, $p > 0.10$), or with higher standard deviation of initial ratings (*Sup_SD*, odds ratio of 1.69, $p > 0.10$). Overall, we find no evidence to support that calibration committees base adjustment decisions on the credibility of ratings as proxied by our variables. Lack of support for this hypothesis indicates that calibration committees fail to consider credibility cues beyond the ratings distributional attributes examined above. Alternatively, lack of support for H2 could be due to inadequate measures of rating credibility. For example, supervisors with wide control spans, instead of benefiting from greater levels of benchmarking information, may be more likely to be overloaded than supervisors with narrow control spans (Aghion and Tirole [1997]) and thus produce less credible ratings.

5.3.3 Hypotheses H3a and H3b

Finally, we examine two structural characteristics associated with the calibration process. First, H3a predicts that calibration committees will be less likely to adjust supervisor ratings of subordinate auditors for whom the hierarchical distance between the subordinate and the committee is greater. We find that, with each increase of one level of hierarchical distance,

calibration committees are approximately 0.6 times as likely to adjust ratings (column 2, odds ratio of 0.638, $p < 0.01$), and any such adjustments are lower in magnitude (column 4, coefficient of -0.022, $p < 0.01$). Thus, we provide evidence in support of H3a by showing that calibration committees defer rating judgments to supervisors when the hierarchical distance between the committee and the auditor being rated is greater.

Second, H3b predicts that calibration committees will be less likely to adjust ratings of a supervisor who serves on their calibration committee We find that calibration committees are approximately 0.6 times as likely to adjust ratings of supervisors on the calibration committee as compared to supervisors who are not on the committee (column 2, odds ratio of 0.678, $p < 0.01$), and any adjustments for supervisors on the committee are lower in magnitude (column 4, coefficient of -0.020, $p < 0.05$). Thus, we provide evidence in support of H3b.

5.3.4 Controlling for Initial Rating

In column 1 of Table 4, we document determinants of supervisor initial ratings. One such determinant was *Hier_Dist*, the variable of interest in the test of H3a. *Hier_Dist* is negatively associated with *Initial_Rating* (column 1, coefficient of -0.144, $p < 0.10$). To rule out the possibility that our results supporting H3a are driven by differences in the initial rating, we repeat our analysis with an additional control for *Initial_Rating* (Table 4, columns 5-7). Inferences regarding H1b, H2, H3a, and H3b are unchanged when including this additional control variable.

5.4 SENSITIVITY ANALYSES: ADJUSTMENTS AFFECTING PAY

As can be seen in Figure 1, Panel A, it is possible for the calibration committee to adjust a rating but for that adjustment to have no effect on subordinate auditor pay. For example, an initial rating of 3 that is adjusted to 3.5 by the calibration committee does not change the number

of bonus shares the auditor would receive (i.e., 2), and hence the auditor's pay will be unaffected. Accordingly, we consider whether our hypothesized effects hold when we redefine our dependent variables to reflect only those adjustments affecting pay. In untabulated analyses, calibration committees made rating adjustments that actually altered the number of bonus shares the auditor received 192 (59%) out of the 327 times an adjustment was made. Of those 192 adjustments, 164 were downward adjustments with more than 90 percent of these resulting in the loss of only one bonus share. The amount by which performance pay was reduced for these 164 auditors averaged approximately $1,000. In contrast, of the 28 times upward adjustments were made that affected pay, all resulted in a gain of only one bonus share and, again, with an average pay impact of approximately $1,000.

Table 5 repeats the analyses presented in Table 4 but with dependent variables reflecting only those changes that affected pay (column 1 is repeated from Table 4 for reference). Inferences based on Table 5 are largely unchanged, with two exceptions. First, we no longer find support for the previous finding that calibration committees are less likely to adjust ratings of supervisors that serve on the committee (column 2, odds ratio of 0.786, $p > 0.10$), nor do any such adjustments differ in magnitude (column 4, coefficient of -0.011, $p > 0.10$). Second, when focusing more narrowly on only those adjustments that affect pay, we find that, consistent with H2, calibration committees are less likely to adjust ratings for supervisors with relatively higher spans of control, one of our three proxies for rating credibility. Specifically, for each increase of one in control span, the committee is about 0.9 times as likely to adjust the supervisor's ratings (column 2, odds ratio of 0.930, $p < 0.01$). Further, any such adjustments are smaller in magnitude (column 4, coefficient of -0.004, $p < 0.01$). These exceptions hold when controlling for supervisor initial ratings (columns 5 and 7).

5.6 SUPPLEMENTAL EVIDENCE OF JUSTICE PERCEPTIONS

As described in Section 2, the biases and inconsistencies that can arise in subjective performance evaluation systems may reduce perceptions of organizational justice. Thus, one of the implied purposes of the calibration process is to improve perceptions of fairness and justice. Prior research documents benefits of organizational justice, in the form of increased employee effort, decreased employee-supervisor conflict, and decreased employee turnover. In these studies, however, it is typically the employees' *perceptions* of justice in the performance evaluation and reward system that is examined. Thus, we conclude our analysis with descriptive evidence of employees' and supervisors' perceptions of organizational justice in our setting.

Auditors received their 2008 ratings and were awarded their bonuses in January 2009. In February 2009, we surveyed auditors to gauge their perceptions regarding the fairness of, and their satisfaction with, the evaluation system. Table 6, Panel A provides descriptive statistics of the survey questions, all of which are measured on a 7-point Likert scale. Auditors, on average, perceived that the outcome of their evaluation was fair, but they were not satisfied with the system itself, and they believed that favoritism was applied on the part of supervisors and/or the calibration committee.[15] Supervisors and program directors, on average, perceived that the outcome and process of auditor evaluations was fair and that favoritism was not an issue. However, they were also dissatisfied with the system.[16]

Table 6, Panel B reports the correlations among these survey questions, a measure of overall auditor performance (*Final_Shares*) and the indicators for downward and upward

[15] The mean *Outcome_Fairness* and *Favoritism* responses were significantly greater than the neutral response of 4 ($t = 4.70$ and $t = 4.25$, respectively, $p < 0.01$); the mean *Satisf_System* response was significantly less than the neutral response ($t = -7.88$, $p < 0.01$); the mean *Process_Fairness* response was not different than the neutral response ($t = -0.30$, $p > 0.10$). All tests are two-tailed.

[16] The mean *Outcome_Fairness* and *Process_Fairness* responses were significantly greater than the neutral response ($t = 4.75$ and $t = 2.47$, respectively, $p < 0.05$); the mean *Satisf_System* and *Favoritism* responses were significantly less than the neutral response ($t = -2.32$ and $t = -2.51$, respectively, $p < 0.05$; all two-tailed tests).

adjustments (*Down_Adj* and *Up_Adj*). One noteworthy observation is the lack of a significant correlation between auditor fairness perceptions and whether their performance pay was adjusted. This provides evidence that supervisors did not share their initial ratings with auditors prior to the calibration process, consistent with formal guidelines prohibiting such dissemination of pre-calibrated ratings.[17]

Among calibration committee members and senior executives, the belief that the calibration process, overall, effectively promotes the fairness and incentive effects of their pay-for-performance evaluation system is strong and widespread. That this system promotes improved auditor performance despite the lack of perceived fairness in the process is supported by prior research which finds that the fairness in outcomes is a more important predictor of job satisfaction and performance than is fairness associated with the process (McFarlin and Sweeney [1992], Lowe and Vodanovich [1995]). Thus, because auditors report a perception of outcome fairness, they will be positively motivated by the evaluation system in place, notwithstanding dissatisfaction with the overall system and perceptions of favoritism.

6. Summary

This study provides the first empirical evidence on the role of calibration committees in subjective performance evaluation systems. In doing so, we document both intended effects, such as improved cross-sectional rating consistency, as well as unintended effects in the form of exacerbated centrality bias. Despite the recognized importance of gaining a better understanding of this practice, there is virtually no empirical evidence on this subject, primarily due to the

[17] Another noteworthy observation is that higher- (lower-) performing auditors (as measured by *Final_Shares*) report higher (lower) levels of perceived fairness and evaluation system satisfaction, and lower (higher) levels of favoritism. Although the organization would have preferred to see higher mean fairness scores, they believed fairness perceptions would be even lower without the system in place. They also interpreted the correlations as evidence that the system is working as intended. Because a primary goal of the performance evaluation system is to retain higher-performing auditors, they are especially interested in these auditors having higher levels of fairness and satisfaction perceptions (and lower levels of favoritism perceptions) as compared to lower-performing auditors.

unavailability of data.

This study also sheds light on the reality that immediate supervisors do not always have exclusive decision rights over the ratings of subordinate employees. Prior archival research on subjective rating biases (e.g., leniency and centrality) examines final, potentially adjusted, subjective ratings. Our results show that, because calibration committees appear to reduce leniency bias, the bias of immediate supervisors may be higher than previous research suggests. Further, our results show an asymmetric response to high ratings versus low ratings. Specifically, while calibration committees mitigate inflated ratings, this could possibly be at the expense of increasing understated ratings. Paradoxically, we also show that input from higher level sources may be more responsible than immediate supervisors for increasing centrality bas. These insights are gleaned only by separately examining immediate supervisor initial ratings and final, adjusted ratings.

Using proprietary data from the internal audit department of a large multinational organization, we examine the calibration process and the determinants of adjustments to initial subjective performance ratings. We show that calibration committees tend to adjust extreme high ratings as well as ratings assigned by supervisors with high mean ratings. This is not the case, however, for extreme low ratings and ratings given by supervisors with low mean ratings. Thus, we document an asymmetric preference of calibration committees related to adjustments in the calibration process. We also show that downward adjustments are more common and that the magnitude of downward adjustments is greater than the magnitude of upward adjustments, which results in a final, adjusted rating distribution that has a lower mean and lower standard deviation relative to the distribution of initial ratings. Thus, to the extent that the initial distribution of subjective ratings exhibits the well-documented leniency and centrality biases, calibration

committees appear to mitigate leniency bias, but exacerbate centrality bias.

Contrary to our prediction, we find no association between calibration committee adjustments and our proxies for the credibility of supervisor ratings. However, we do find two structural characteristics of the calibration process that are associated with calibration committee adjustments. First, we find that committee adjustments are decreasing in the hierarchical distance between the committee and the subordinate auditor being rated, consistent with the collocation of decision-making authority with the individuals possessing the relevant knowledge, and, second, that adjustments are less likely when the rating supervisor serves on the committee.

This study is subject to the usual generalizability caveats of a field study. In this organization, work is complex, supervisors generally have narrow control spans, bonus pool payouts are fixed, etc. Therefore, the results may not generalize to settings with different characteristics. Nevertheless, while the data come from one organization, the hypotheses are not specific to this one organization; rather, the hypotheses generalize to organizational settings in which subjective ratings are used and a calibration process is employed.

This study suggests several avenues for future research. First, it is unclear whether using a calibration process is the most effective and efficient mechanism by which to mitigate supervisor bias and promote cross-sectional consistency in ratings. Organizations can also rotate supervisors, establish methods for keeping supervisors responsible for promoted subordinates, more closely tie supervisor rewards to employee performance, implement appeals procedures, have employees evaluate supervisors, use a type of forced-distribution or ranking format, etc. (Prendergast and Topel [1993]). Second, the ramifications of the strategic interactions between the supervisors making the initial ratings and the committee members calibrating those ratings is beyond the scope of this study due to the short time series of our data. Organizational dynamics

across hierarchical levels are possibly one of the least studied and least understood aspects of subjective performance evaluation. There is ample opportunity, data permitting, for future research to examine strategic actions over time by both parties as they seek to advance their own interests. In summary, to our knowledge, this study presents the first empirical evidence of the role of calibration committees within an organization's performance evaluation system. While it provides novel insights, it also raises many new and interesting lines of inquiry.

REFERENCES

AGHION, P., and J. TIROLE. "Formal and real authority in organizations." *Journal of Political Economy* 105 (1997): 1–29.

ANDERSON, S. W., H. DEKKER, K. L. SEDATOLE, and E. WIERSMA. "Field evidence of bias in subjective ratings," Unpublished paper, University of California at Davis, 2014. Working paper.

ARVEY, R. D., and K. R. MURPHY. "Performance ratings in work settings." *Annual Review of Psychology* 49 (1998): 141–68.

BAIMAN, S. and M. V. RAJAN. "The informational advantages of discretionary bonus schemes." *The Accounting Review* 70 (1995): 557–79.

BEAULIEU, P. R. "The effects of judgments of new clients' integrity upon risk judgments, audit evidence, and fees." *Auditing: A Journal of Practice & Theory* 20 (2001): 85-99.

BEAULIEU, P. R. "Commercial lenders' use of accounting information in interaction with source credibility." *Contemporary Accounting Research* 10 (1994): 557-585.

BIRNBAUM, M. H., and S. E. STENGER. "Source credibility in social judgment: Bias, expertise, and the judge's point of view." *Journal of Personality and Social Psychology* 37 (1979): 48-74.

BJERKE, D., CLEVELAND, J., MORRISON, R., and WILSON, W. "Officer Fitness report evaluation study." Navy Personnel Research and Development Center report no. TR 88-4. Navy Personnel Research and Development Center, 1987.

BOL, J. C. "Subjectivity in compensation contracting." *Journal of Accounting Literature* 27 (2008): 1-24.

BOL, J. "The determinants and performance effects of managers' performance evaluation biases." *The Accounting Review* 86 (2011): 1549-75.

BRETZ, R. D., G. T. MILKOVICH and W. READ. "The Current State of Performance evaluation Research and Practice: Concerns, Directions, and Implications." *Journal of Management* 18 (1992): 321-52.

CAMPBELL, D. "Nonfinancial Performance Measures and Promotion-Based Incentives." *Journal of Accounting Research* 46 (2008): 297-332.

CARDINAELS, E., and E. LABRO. "On the determinants of measurement error in time-driven costing." *The Accounting Review* 83 (2008): 735-56.

CHENHALL, R. H. "Management control system design within its organizational context: Findings from contingency-based research and directions for the future." *Accounting, Organizations and Society* 28 (2003): 127-68.

COLETTI, A. L., K. L. SEDATOLE, and K. L. TOWRY. "The effect of control systems on trust and cooperation in collaborations." *The Accounting Review*, 80 (2005): 477–500.

COLQUITT, J. A. "On the dimensionality of organizational justice: A construct validation of a measure." *Journal of Applied Psychology* 86 (2001): 386-400.

COLQUITT, J. A., D. CONLON, M. J. WESSON, C. O. L. H. PORTER, and K. Y. NG. "Justice at the Millennium: A meta-analytic review of 25 years of organizational justice research." *Journal of Applied Psychology* 86 (2001):425-445.

COLQUITT, J. A., B. A. SCOTT, J. B. RODELL, D. M. LONG, C. P. ZAPATA, D. E. CONLON, and M. J. WESSON. "Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives." *Journal of Applied Psychology* 98 (2013): 199-236.

DEMERÉ, B. W., R. KRISHNAN, K. L. SEDATOLE, and A. WOODS. "Do the incentive effects of relative performance evaluation vary with the *ex ante* probability of promotion?" Working paper, Michigan State University, 2015.

DHOLAKIA, R. R., and B. STERNTHAL. "Highly credible sources: Persuasive facilitators or persuasive liabilities?" *Journal of Consumer Research* 3 (1977): 223-232.

FISHER, J. G., L. A. MAINES, S. A. PEFFER, and G. B. SPRINKLE. "An experimental investigation of employer discretion in employee performance evaluation and compensation." *The Accounting Review* 80 (2005): 563–83.

GIEBE, T., and O. GÜRTLER. "Optimal contracts for lenient supervisors." *Journal of Economic*

Behavior and Organization 81 (2012): 403–420.
GIBBS, M., K. A. MERCHANT, W. A. VAN DER STEDE, and M. E. VARGUS. "Determinants and effects of subjectivity in incentives." *The Accounting Review* 79 (2004): 409–36.
GOLMAN, R. and S. BHATIA. "Performance evaluation inflation and compression." *Accounting, Organizations and Society* 37 (2012) 534–543.
GOW, I. D., G. ORMAZABAL, and D. J. TAYLOR. "Correcting for Cross-sectional and time-series dependence in accounting research." *The Accounting Review* 85 (2010): 483–512.
GRANT, R. M. "Toward a knowledge-based theory of the firm. *Strategic Management Journal* 17 (1996): 109-122.
GROTE, D. *Forced Ranking*. Boston: Harvard Business School Press, 2005.
HARRIS, M. M. "Rater motivation in the performance evaluation context: A theoretical framework." *Journal of Management* 20 (1994): 737-756.
HÖPPE, F. and F. MOERS. "The choice of different types of subjectivity in CEO annual bonus contracts." *The Accounting Review* 86 (2011): 2023-2046.
HOPWOOD, A. G. "An empirical study of the role of accounting data in performance evaluation." *Journal of Accounting Research* 10 (1972): 156–82.
HUBER, V. L. "Comparison of the effects of specific and general performance standards on performance appraisal decisions." *Decision Sciences* 20 (1989): 545-57.
ITTNER, C. D., D. F. LARCKER, and M. MEYER. "Subjectivity and the weighting of performance measures: Evidence from a balanced scorecard." *The Accounting Review* 78 (2003): 725–58.
JENSEN, M. C., and W. H. MECKLING. "Specific and General Knowledge and Organizational Structure." In *Contract Economics*, edited by L. Werin and H. Wijkander. Oxford, U.K.: Blackwell Publishers, 1992.
LIBERTI, J. M., and A. R. MIAN. "Estimating the effect of hierarchies on information use." *The Review of Financial Studies* 22 (2009): 4057-90.
LOWE, R. H. and S. J. VODANOVICH. "A Field Study of Distributive and Procedural Justice as Predictors of Satisfaction and Organizational Commitment." *Journal of Business and Psychology* 10 (1995): 99-114.
MACLEOD, B. W. "Optimal contracting with subjective evaluation." *The American Economic Review* 93 (2003): 216–40.
MCDONALD, P. "Staffing today's internal audit function: audit executives need a realistic strategy for obtaining top talent to handle growing demands." *Internal Auditor* 60 (2003): 46-51.
MCFARLIN, D. B. and P. D. SWEENEY. "Distributive and Procedural Justice as Predictors of Satisfaction with Personal and Organizational Outcomes." *The Academy of Management Journal* 35 (1992): 626-637.
MCGREGOR, J. "The struggle to measure performance." *Businessweek* Issue 3966 (2006): 26-8.
MERCHANT, K. A. *Rewarding Results*. Boston: Harvard Business School Press, 1989.
MILKOVICH, G. T., and A. K. WIGDOR. *Pay for Performance: Evaluating Performance evaluation and Merit Pay*. Washington, D.C.: National Academy Press, 1991.
MOERS, F. "Discretion and bias in performance evaluation: The impact of diversity and subjectivity." *Accounting, Organizations and Society* 30 (2005): 67–80.
MOERS, F. "Performance measure properties and delegation." *The Accounting Review* 81 (2006): 897–924.
MOHRMAN, A. M., S. M. RESNICK-WEST, and E. E. LAWLER. *Designing Performance evaluation Systems: Aligning Appraisals and Organizational Realities*. San Francisco: Jossey-Bass, 1989.
MURPHY, K. R., and J. N. CLEVELAND. *Performance Appraisal: An Organizational Perspective*. Boston: Allyn and Bacon, 1991.
PETERSEN, M. A. "Estimating standard errors in finance panel data sets: Comparing approaches." *Review of Financial Studies* 22 (2009): 435-80.
PORNPITAKPAN, C. "The persuasiveness of source credibility: A critical review of five decades' evidence." *Journal of Applied Social Psychology* 34 (2004): 243-281.

PRENDERGAST, C. "The provision of incentives in firms." *Journal of Economic Literature* 37 (1999): 7–63.
PRENDERGAST, C. "The tenuous trade-off between risk and incentives." *Journal of Political Economy* 110 (2002): 1071–1102.
PRENDERGAST, C., and R. H. TOPEL. "Discretion and bias in performance evaluation." *European Economic Review* 37 (1993): 355–65.
PRENDERGAST, C., & TOPEL, R. H. Favoritism in organizations. *Journal of Political Economy,* 104 (1996): 958–978.
RAJAN, M. V., and S. REICHELSTEIN. "Subjective performance indicators and discretionary bonus pools." *Journal of Accounting Research* 44 (2006): 585–618.
RAJAN, M. V., and S. REICHELSTEIN. "Objective versus Subjective Indicators of Managerial Performance." *The Accounting Review* 84 (2009): 209–37.
REICHERT, B., and A. WOODS. "The effect of orientation on employee performance." Working paper, College of William and Mary, 2015.
RISHER, H. "Getting performance management on track." *Compensation & Benefits Review* 43 (2011): 273-281.
RISHER, H. "Reward management depends increasingly on procedural justice." *Compensation & Benefits Review* 46 (2014): 135-138.
SAAL, F. E. and F. J. LANDY. "The mixed standard rating scale: An evaluation." *Organizational Behavior and Human Performance* 18 (1977):19-35.
SCHAPPE, P. "Understanding Employee Job Satisfaction: The Importance of Procedural and Distributive Justice." *Journal of Business and Psychology* 12 (1998): 493-503
SCHNEIDER, D. E., and A. G. BAYROFF. "The relationship between rater characteristics and validity of ratings." *Journal of Applied Psychology*, 37 (1953): 278–80.
SEDATOLE, K., A. SWANEY, and A. WOODS. "The implicit incentive effects of horizontal monitoring and team member dependence on individual performance." *Contemporary Accounting Research*, Forthcoming.
SOCIETY FOR HUMAN RESOURCE MANAGEMENT. *SHRM Poll: Performance Management and Other Workplace Practices*, 2011. Available from http://www.shrm.org/research/surveyfindings/articles/pages/performancemanagementandotherworkplacepracticesshrmpollfinal.aspx
SPENCE, J. R. and L. M. KEEPING. "The impact of non-performance information on ratings of job performance: A policy-capturing approach." *Journal of Organizational Behavior* 31 (2010): 587–608.
TAN, H., and K. JAMAL. "Do auditors objectively evaluate their subordinates' work?" *The Accounting Review* 76 (2001): 99–110.
WOODS, A. "Subjective adjustments to objective performance measures: the influence of prior performance." *Accounting, Organizations and Society* 37 (2012): 403-425.

APPENDIX
Variable Definitions

Variable Name		Definition
Initial_Rating	=	the initial rating made by a supervisor.
Adjusted_Rating	=	final (committee-adjusted) rating.
Adjustment	=	indicator variable that equals 1 for ratings adjusted by the calibration committee (either up or down), and 0 otherwise.
Adj_Magnitude	=	the magnitude of the committee adjustment, *Adjusted_Rating* less *Initial_Rating*.
Adj_Abs_Magnitude	=	the absolute magnitude of the committee adjustment (*Adj_Magnitude*).
Adj_Direction	=	multilevel categorical indicator variable that equals "DOWN" for ratings adjusted downward by the calibration committee, "NONE" for ratings not adjusted, and "UP" for ratings adjusted upward by the committee.
Extreme_High	=	indicator variable that equals 1 for ratings that are above 4.00 (resulting in 4 or more bonus shares), and 0 otherwise.
Extreme_Low	=	indicator variable that equals 1 for ratings that are below 3.00 (resulting in 0 or 1 bonus shares), and 0 otherwise.
High_Sup_Avg	=	indicator variable that equals 1 for supervisors whose mean initial rating (in a given pool and year) is above 4.00, and 0 otherwise.
Low_Sup_Avg	=	indicator variable that equals 1 for supervisors whose mean initial rating (in a given pool and year) is below 3.00, and 0 otherwise.
Sup_Perf	=	supervisor's overall performance.
Sup_Span	=	the number of auditors that report directly to a supervisor.
Sup_SD	=	the standard deviation of the supervisor's subordinate auditor ratings being evaluated by a given calibration committee.
Hier_Dist	=	the hierarchical distance between the calibration committee and the subordinate auditor whose rating is being evaluated for adjustment.
Sup_Member	=	indicator variable that equals 1 if an auditor's first-level supervisor or second-level supervisor is a member of the calibration committee for that auditor's bonus pool, and 0 otherwise.
Certification	=	indicator variable that equals 1 if an auditor has a professional certification (e.g., CPA, CIA), and 0 otherwise.
Ineligible	=	indicator variable that equals 1 if an auditor is ineligible for promotion in a given year, and 0 otherwise.
Hier_Level	=	the subordinate auditor's level in the organizational hierarchy, where 1 = Junior Auditor and 5 = Program Director (see Figure 1).
Year [X]	=	indicator variable that equals 1 for a given evaluation year (Year 2007, 2008, or 2009), and 0 otherwise.
Committee [X]	=	indicator variable that equals 1 for a given calibration committee (1-12), and 0 otherwise.

FIGURE 1
Organization hierarchy, calibration committee structure, and bonus pool shares

Panel A: Relation of performance level to bonus pool shares

Overall Final Rating (i.e., adjusted)		Performance Level		# of Bonus Pool Shares
1.00—2.50	=	1	=	0
2.51—2.99	=	2	=	1
3.00—3.50	=	3	=	2
3.51—4.00	=	4	=	3
4.01—4.50	=	5	=	4
4.51—4.75	=	6	=	5
4.76—5.00	=	7	=	6

Panel B: Organization hierarchy and calibration committee structure

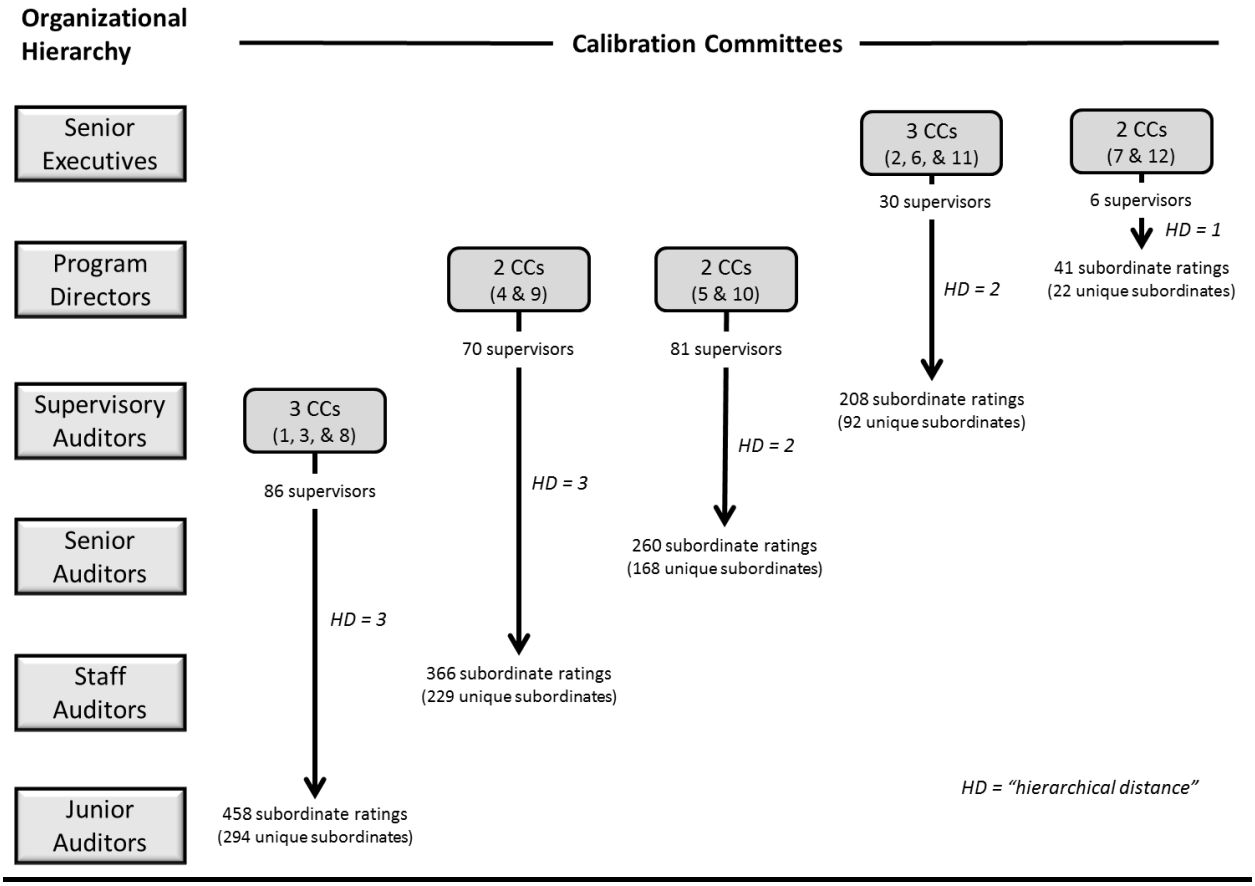
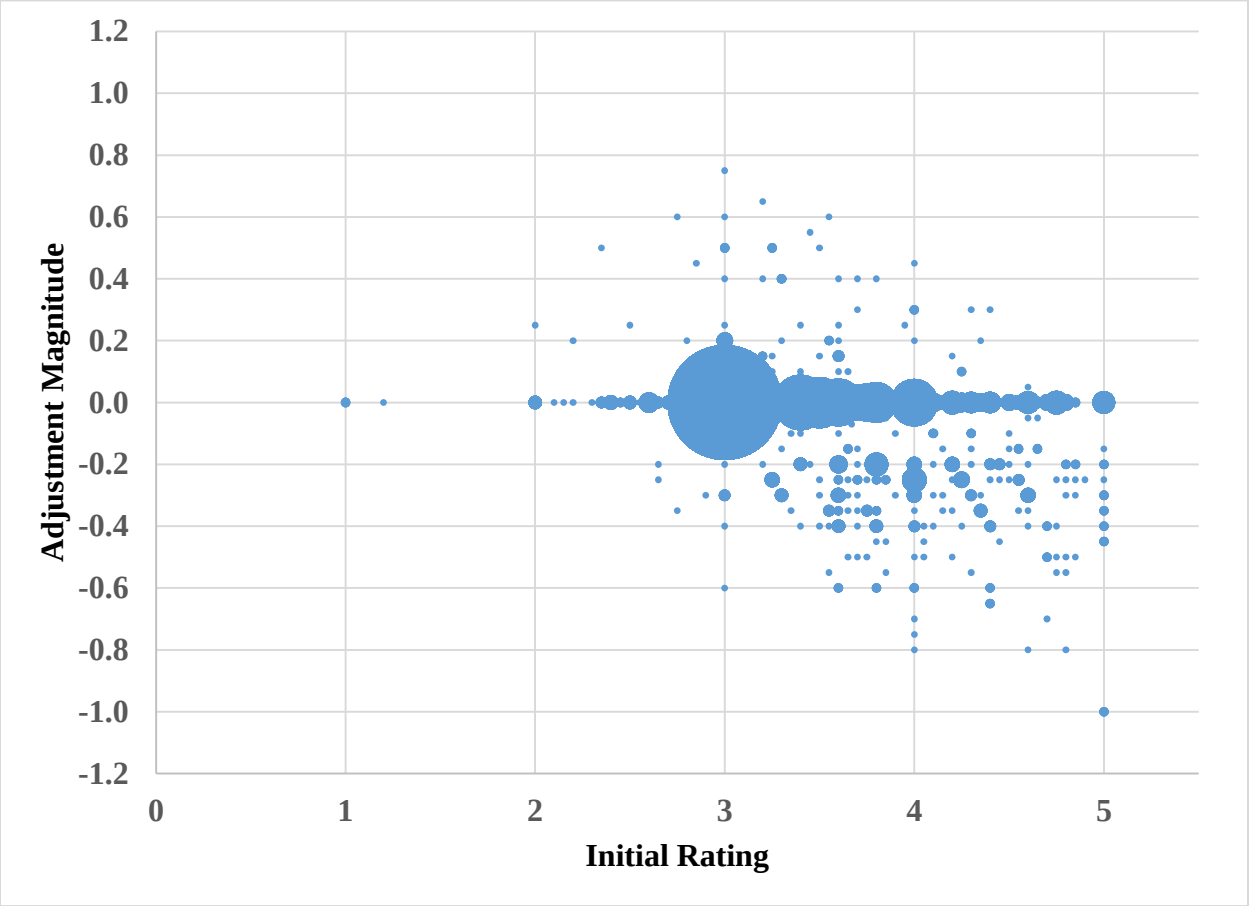


FIGURE 2
Calibration committee adjustment plotted against the initial rating



The figure graphically depicts the distribution of calibration adjustments as a function of initial ratings. The size of the bubbles corresponds to the density of initial rating-adjustment magnitude combinations.

TABLE 1
Descriptive statistics

Panel A: Initial vs. adjusted ratings by calibration committee

Calibration Committee	# Subordinate Ratings	# Supervisors	Initial Rating Mean (st dev)	Adjusted Rating Mean (st dev)	Difference Mean t-stat (st dev F-stat)
1	143	46	3.564 (0.580)	3.534 (0.540)	1.995** (1.153)
2	69	24	3.814 (0.606)	3.707 (0.555)	4.771*** (1.195)
3	146	56	3.428 (0.558)	3.425 (0.556)	0.250 (1.006)
4	198	61	3.567 (0.583)	3.500 (0.496)	6.256*** (1.175)
5	120	61	3.567 (0.583)	3.500 (0.496)	4.230*** (1.383*)
6	69	25	3.856 (0.533)	3.800 (0.508)	3.485*** (1.098)
7	21	5	3.869 (0.542)	3.883 (0.498)	-0.373 (1.185)
8	169	57	3.491 (0.619)	3.466 (0.583)	2.281** (1.126)
9	168	53	3.410 (0.579)	3.373 (0.568)	3.390*** (1.041)
10	140	66	3.509 (0.548)	3.453 (0.497)	4.391*** (1.215)
11	70	25	3.880 (0.611)	3.816 (0.542)	2.714*** (1.269)
12	20	5	3.990 (0.423)	3.980 (0.438)	0.233 (0.930)
Total	1,333 (686 unique)	484 (110 unique)	3.559 (0.590)	3.511 (0.552)	10.541*** (1.141**)

This table provides means and standard deviations of initial and adjusted ratings by calibration committee.

Panel B: Frequency distribution of initial ratings by calibration committee

Initial Rating	1.00 – 2.50	2.51 – 2.99	3.00 – 3.50	3.51 – 4.00	4.01 – 4.50	4.51 – 4.75	4.76 – 5.00	Subordinate
Bonus Shares	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Ratings**
1	2	1	76	37	15	3	9	143
2	1	4	16	24	10	10	4	69
3	3	3	91	30	10	6	3	146
4	6	9	81	80	16	3	3	198
5	1	3	67	26	12	7	4	120
6	-	-	23	22	15	7	2	69
7	-	1	2	11	3	3	1	21
8	4	3	90	48	10	6	8	169
9	8	9	82	48	12	7	2	168
10	3	4	78	33	15	1	6	140
11	1	-	23	20	12	10	4	70
12	-	-	2	11	4	3	-	20
Total	29	37	631	390	134	66	46	1,333

This table provides frequencies of initial bonus pool shares (prior to committee adjustment) by calibration committee.

Panel C: Frequency distribution of adjustments by direction and calibration committee

	Initial Rating	1.00 – 2.50	2.51 – 2.99	3.00 – 3.50	3.51 – 4.00	4.01 – 4.50	4.51 – 4.75	4.76 – 5.00	Total by	Total	Subordinate	%
	Bonus Shares	0	1	2	3	4	5	6	direction		ratings	Adjusted
1	DOWN			3	12	8		4	27			
	UP	1		8	2		1		12	39	143	27%
2	DOWN			2	11	5	8	3	29			
	UP		1		1				2	31	69	45%
3	DOWN			8	1	2			11			
	UP	1		4	4				9	20	146	14%
4	DOWN			6	39	7	2	2	56			
	UP		1	4	3				8	64	198	32%
5	DOWN			3	8	8	4	4	27			
	UP	1		1					2	29	120	24%
6	DOWN				4	5	2	1	12			
	UP					2			2	14	69	20%
7	DOWN				2			1	3			
	UP		1		2	1			4	7	21	33%
8	DOWN				4	5	2	4	15			
	UP			1	1				2	17	169	10%
9	DOWN		3	4	13	5	1		26			
	UP	1		5	4	1			11	37	168	22%
10	DOWN		1	5	11	13	1	4	35			
	UP			5					5	40	140	29%
11	DOWN				6	1	6	4	17			
	UP			2	2				4	21	70	30%
12	DOWN			1	1		2		4			
	UP				2	2			4	8	20	40%
Total	DOWN	0	4	32	112	59	28	27	262			20%
	UP	4	3	30	21	6	1	0	65			5%
	Total	4	7	62	133	65	29	27	327	327	1,333	25%

This table provides number of ratings adjusted upward and downward by initial bonus shares and calibration committee.

TABLE 2
Variable descriptive statistics and correlations

Panel A: Descriptive statistics

Variable	Obs	Mean	Std.Dev.	Min	Max
Initial_Rating	1,333	3.559	0.590	1	5
Adjusted_Rating	1,333	3.511	0.552	1	5
Adjustment (0/1)	1,333	0.245	0.430	0	1
Adj_Magnitude	1,333	-0.047	0.163	-1	0.75
Adj_Abs_Magnitude	1,333	0.075	0.153	0	1
Extreme_High (0/1)	1,333	0.185	0.388	0	1
Extreme_Low (0/1)	1,333	0.050	0.217	0	1
High_Sup_Avg (0/1)	1,333	0.117	0.322	0	1
Low_Sup_Avg (0/1)	1,333	0.035	0.183	0	1
Sup_Perf	1,333	3.839	0.530	2.3	5
Sup_Span	1,333	7.174	2.955	1	15
Sup_SD	1,333	0.357	0.240	0.000	1.500
Hier_Dist	1,333	2.587	0.551	1	3
Sup_Member (0/1)	1,333	0.310	0.463	0	1
Certification (0/1)	1,333	0.272	0.445	0	1
Ineligible (0/1)	1,333	0.131	0.337	0	1
Hier_Level	1,333	2.256	1.173	1	5
Year 2007	1,333	0.159	0.366	0	1
Year 2008	1,333	0.416	0.493	0	1
Year 2009	1,333	0.425	0.495	0	1
Committee					
1	1,333	0.107	0.310	0	1
2	1,333	0.052	0.222	0	1
3	1,333	0.110	0.312	0	1
4	1,333	0.149	0.356	0	1
5	1,333	0.090	0.286	0	1
6	1,333	0.052	0.222	0	1
7	1,333	0.016	0.125	0	1
8	1,333	0.127	0.333	0	1
9	1,333	0.126	0.332	0	1
10	1,333	0.105	0.307	0	1
11	1,333	0.053	0.223	0	1
12	1,333	0.015	0.122	0	1

This table provides descriptive statistics for all variables. See Panel B for variable descriptions.

Panel B: Bivariate correlations

	1	2	3	4	5	6	7	8	9	10	11	12	13
1 *Initial_Rating*	**1.000**												
2 *Adjusted_Rating*	**0.961**	**1.000**											
3 *Adjustment*	**0.335**	**0.208**	**1.000**										
4 *Adj_Magnitude*	**-0.361**	**-0.090**	**-0.507**	**1.000**									
5 *Adj_Abs_Magnitude*	**0.314**	**0.159**	**0.859**	**-0.593**	**1.000**								
6 *Extreme_High*	**0.759**	**0.726**	**0.273**	**-0.287**	**0.252**	**1.000**							
7 *Extreme_Low*	**-0.419**	**-0.419**	-0.042	**0.095**	-0.031	**-0.109**	**1.000**						
8 *High_Sup_Avg*	**0.429**	**0.386**	**0.248**	**-0.243**	**0.226**	**0.410**	**-0.083**	**1.000**					
9 *Low_Sup_Avg*	**-0.285**	**-0.276**	-0.041	**0.096**	**-0.048**	**-0.090**	**0.468**	**-0.069**	**1.000**				
10 *Sup_Perf*	**0.187**	**0.164**	0.041	**-0.121**	0.041	**0.181**	-0.043	**0.193**	**-0.065**	**1.000**			
11 *Sup_Span*	**-0.076**	**-0.070**	-0.051	0.036	**-0.067**	**-0.056**	-0.021	**-0.168**	-0.021	**0.060**	**1.000**		
12 *Sup_SD*	**0.167**	**0.146**	**0.083**	**-0.111**	**0.105**	**0.215**	**0.084**	**0.078**	-0.007	**0.191**	**0.129**	**1.000**	
13 *Hier_Dist*	**-0.196**	**-0.196**	**-0.095**	0.046	**-0.075**	**-0.160**	0.052	**-0.134**	**0.112**	**-0.061**	**0.137**	-0.026	**1.000**
14 *Sup_Member*	**0.137**	**0.140**	0.021	-0.020	0.019	**0.125**	-0.018	**0.104**	**-0.056**	0.035	**-0.188**	**0.135**	**-0.405**

This table provides correlations for key variables**. Bolded** Pearson correlations are significant at the 5% level (two-tailed).
Initial_Rating is the initial rating made by a supervisor. *Adjusted_Rating* is the final (committee-adjusted) rating. *Adjustment* is an indicator variable that equals 1 for ratings adjusted by the calibration committee (either up or down), and 0 otherwise. *Adj_Magnitude* is the magnitude of the committee adjustment, *Adjusted_Rating* less *Initial_Rating*. *Adj_Abs_Magnitude* is the absolute magnitude of the committee adjustment. *Extreme_High* (*Extreme_Low*) is an indicator variable that equals 1 for ratings that are above 4.00 (below 3.00), and 0 otherwise. *High_Sup_Avg* (*Low_Sup_Avg*) is an indicator variable that equals 1 for supervisors whose mean initial rating (in a given year) is above 4.00 (below 3.00), and 0 otherwise. *Sup_ Perf* is the supervisor's overall performance. *Sup_Span* is the number of auditors that report directly to a supervisor. Sup_SD is the standard deviation of the supervisor's subordinate auditor ratings being evaluated by a given calibration committee. Hier_Dist is the hierarchical distance between the calibration committee and the subordinate auditor whose rating is being evaluated for adjustment. *Sup_Member* is an indicator variable equals 1 if an auditor's first-level supervisor or second-level supervisor is a member of the calibration committee, and 0 otherwise. *Certification* is indicator variable that equals 1 if an auditor has a professional certification, and 0 otherwise. *Ineligible* is indicator variable that equals 1 if an auditor is ineligible for promotion in a given year, and 0 otherwise. *Hier_Level* is the subordinate auditor's level in the organizational hierarchy, where 1 = Junior Auditor and 5 = Program Director (see Figure 1). *Year [X]* are year indicator variables. *Committee [X]* are calibration committee indicator variables.

Table 3
Variable descriptive statistics by direction of calibration committee adjustments

	***Adj_Direction* = DOWN N=262**		***Adj_Direction* = NONE N = 1,006**		***Adj_Direction* = UP N = 65**	
	Mean	**Std. Dev.**	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Initial_Rating	4.031	0.524	3.446	0.548	3.398	0.524
Adj_Magnitude	-0.310	0.155	0.000	0.000	0.283	0.170
Adjusted_Rating	3.721	0.521	3.446	0.548	3.682	0.508
Extreme_High	0.435	0.497	0.124	0.330	0.108	0.312
Extreme_Low	0.015	0.123	0.055	0.227	0.108	0.312
High_Sup_Avg	0.302	0.460	0.072	0.258	0.077	0.269
Low_Sup_Avg	0.000	0.000	0.039	0.193	0.108	0.312
Sup_Perf	3.933	0.525	3.826	0.520	3.652	0.627
Sup_Span	6.908	3.006	7.260	2.931	6.908	3.086
Sup_SD	0.405	0.229	0.345	0.243	0.335	0.222
Hier_Dist	2.489	0.552	2.617	0.537	2.523	0.709
Sup_Member	0.332	0.472	0.304	0.460	0.308	0.465
Certification	0.332	0.472	0.258	0.438	0.246	0.434
Ineligible	0.065	0.247	0.149	0.356	0.108	0.312
Hier_Level	2.557	1.122	2.170	1.157	2.369	1.398
Year 2007	0.214	0.411	0.141	0.348	0.215	0.414
Year 2008	0.416	0.494	0.417	0.493	0.385	0.490
Year 2009	0.370	0.484	0.441	0.497	0.400	0.494

This table provides descriptive statistics for key variables by levels of *Adj_Direction* (DOWN, NONE, or UP).
Initial_Rating is the initial rating made by a supervisor. *Adjusted_Rating* is the final (committee-adjusted) rating. *Adjustment* is an indicator variable that equals 1 for ratings adjusted by the calibration committee (either up or down), and 0 otherwise. *Adj_Magnitude* is the magnitude of the committee adjustment, *Adjusted_Rating* less *Initial_Rating*. *Adj_Abs_Magnitude* is the absolute magnitude of the committee adjustment. *Extreme_High* (*Extreme_Low*) is an indicator variable that equals 1 for ratings that are above 4.00 (below 3.00), and 0 otherwise. *High_Sup_Avg* (*Low_Sup_Avg*) is an indicator variable that equals 1 for supervisors whose mean initial rating (in a given year) is above 4.00 (below 3.00), and 0 otherwise. *Sup_ Perf* is the supervisor's overall performance. *Sup_Span* is the number of auditors that report directly to a supervisor. Sup_SD is the standard deviation of the supervisor's subordinate auditor ratings being evaluated by a given calibration committee. Hier_Dist is the hierarchical distance between the calibration committee and the subordinate auditor whose rating is being evaluated for adjustment. *Sup_Member* is an indicator variable equals 1 if an auditor's first-level supervisor or second-level supervisor is a member of the calibration committee, and 0 otherwise. *Certification* is indicator variable that equals 1 if an auditor has a professional certification, and 0 otherwise. *Ineligible* is indicator variable that equals 1 if an auditor is ineligible for promotion in a given year, and 0 otherwise. *Hier_Level* is the subordinate auditor's level in the organizational hierarchy, where 1 = Junior Auditor and 5 = Program Director (see Figure 1). *Year [X]* are year indicator variables.

Table 4
Tests of hypotheses

	OLS		Logit	Multinomial Logit	OLS		Logit	Multinomial Logit	OLS
	(1)		(2)	(3)	(4)		(5)	(6)	(7)
DV:	Initial_ Rating		Adjustment (Adj vs None)	Adjustment_ Direction (Up vs. Down)	Adj_Abs_ Magnitude		Adjustment (Adj vs None)	Adjustment_ Direction (Up vs. Down)	Adj_Abs_ Magnitude
Variable	**Coeff.**	Predicted Odds Ratio	**Odds Ratio**	**Odds Ratio**	**Coeff.**	Predicted Odds Ratio	**Odds Ratio**	**Odds Ratio**	**Coeff.**
Extreme_High		H1a: >1	3.051 ***	0.219 ***	0.071 ***	H1a: >1	[a]	[a]	[a]
Extreme_Low		H1a: >1	0.711	3.302	-0.009	H1a: >1	[a]	[a]	[a]
High_Sup_Avg		H1b: >1	3.071 ***	0.447	0.070 ***	H1b: >1	2.225 **	0.660	0.054 **
Low_Sup_Avg		H1b: >1	1.163	> 100 ***	-0.009	H1b: >1	2.281 **	>100 ***	0.034 *
Sup_Perf	0.176 ***	H2: <1	0.864	0.498	-0.008	H2: <1	0.823	0.528	-0.010
Sup_Span	-0.008	H2: <1	0.984	1.026	-0.003	H2: <1	0.989	1.016	-0.003
Sup_SD	0.313 ***	H2: <1	1.690	0.387	0.049 *	H2: <1	1.349	0.600	0.048 **
Hier_Dist	-0.144 *	H3a: <1	0.638 ***	0.506 **	-0.022 ***	H3a: <1	0.752 ***	0.409 ***	-0.012 **
Sup_Member	0.035	H3b: <1	0.678 ***	0.866	-0.020 **	H3b: <1	0.676 ***	0.873	-0.019 **
Certification	0.090								
Ineligible	-0.200 ***								
Hier_Level	-0.018								
Initial_Rating							3.750 ***	0.158 ***	0.071 ***
Intercept	3.326 ***		1.464	89.979	0.155 **		0.012 ***	>100 ***	-0.107
Committee Indicators	No		Yes	Yes	Yes		Yes	Yes	Yes
Year Indicators	Yes		No	No	No		No	No	No
Observations	1,333		1,333	1,333	1,333		1,333	1,333	1,333
Supervisor Clusters	110		12	12	12		12	12	12
R^2/Pseudo R^2	0.1108		0.1210	0.1524	0.1133		0.1546	0.1904	0.1404

*** p<0.01, ** p<0.05, * p<0.10 indicate two-tailed (one-tailed for predicted signs) significance using robust standard errors clustered by calibration committee.

We report the OLS estimation of *Initial_Rating* (column 1) as a reference. The test of hypotheses are reported in columns 2-7 with the logistic estimation of *Adjustment* (columns 2 and 5), the multinomial logistic estimation of *Adj_Direction* (columns 3 and 6, using DOWN as the base category), and the OLS estimation of *Adj_Abs_Magnitude* (columns 4 and 7) as a function of our variables of interest and control variables. Regression coefficients are reported in columns 1, 4, and 7, and odds ratios are reported in columns 2, 3, 5, and 6. Columns 5-6 repeat the analysis of columns 2-4 but with an additional control for *Initial_Rating*. Note that there were no downward adjustments for supervisors giving low mean ratings. See the Appendix for variable descriptions.

[a] Omitted because of collinearity with additional, *Initial_Rating* variable.

Table 5

Sensitivity analysis: Determinants of calibration committee adjustments that affect pay

	OLS		Logit	Multinomial Logit	OLS		Logit	Multinomial Logit	OLS
	(1)		(2)	(3)	(4)		(5)	(6)	(7)
DV:	Initial_ Rating		Adjustment (Adj vs None)	Adjustment_ Direction (Up vs. Down)	Adj_Abs_ Magnitude		Adjustment (Adj vs None)	Adjustment_ Direction (Up vs. Down)	Adj_Abs_ Magnitude
Variable	Coeff.	Predicted Odds Ratio	Odds Ratio	Odds Ratio	Coeff.	Predicted Odds Ratio	Odds Ratio	Odds Ratio	Coeff.
Extreme_High		H1a: >1	5.682 ***	0.149 *	0.088 ***	H1a: >1	[a]	[a]	[a]
Extreme_Low		H1a: >1	1.454	7.760	0.007	H1a: >1	[a]	[a]	[a]
High_Sup_Avg		H1b: >1	1.780 **	0.342	0.038 **	H1b: >1	1.445	0.394	0.033 *
Low_Sup_Avg		H1b: >1	0.430	> 100 ***	-0.020	H1b: >1	1.253	>100 ***	0.025
Sup_Perf	0.176 ***	H2: <1	0.902	0.278 *	-0.005	H2: <1	0.901	0.282 *	-0.005
Sup_Span	-0.008	H2: <1	0.930 ***	0.993	-0.004 ***	H2: <1	0.937 ***	0.969	-0.004 ***
Sup_SD	0.313 ***	H2: <1	1.509	0.321	0.030	H2: <1	1.320	0.698	0.036 **
Hier_Dist	-0.144 *	H3a: <1	0.627 ***	0.215 ***	-0.021 ***	H3a: <1	0.717 ***	0.184 ***	-0.014 ***
Sup_Member	0.035	H3b: <1	0.786	0.258 ***	-0.011	H3b: <1	0.788	0.251 ***	-0.010
Certification	0.090								
Ineligible	-0.200 ***								
Hier_Level	-0.018								
Initial_Rating							4.660 ***	0.150 ***	0.065 ***
Intercept	3.326 ***		0.675	>100 **	0.127 **		0.003 ***	>100 ***	-0.109
Committee Indicators			Yes	Yes	Yes		Yes	Yes	Yes
Year Indicators	Yes								
Observations	1,333		1,333	1,333	1,333		1,333	1,333	1,333
Supervisor Clusters	110		12	12	12		12	12	12
R^2/Pseudo R^2	0.1108		0.1629	0.1966	0.1130		0.1738	0.2074	0.1189

*** p<0.01, ** p<0.05, * p<0.10 indicate two-tailed (one-tailed for predicted signs) significance using robust standard errors clustered by calibration committee.

We report the OLS estimation of *Initial_Rating* (column 1) as a reference. The test of hypotheses are reported in columns 2-7 with the logistic estimation of *Adjustment* (columns 2 and 5), the multinomial logistic estimation of *Adj_Direction* (columns 3 and 6, using DOWN as the base category), and the OLS estimation of *Adj_Abs_Magnitude* (columns 4 and 7) as a function of our variables of interest and control variables. Regression coefficients are reported in columns 1, 4, and 7, and odds ratios are reported in columns 2, 3, 5, and 6. Columns 5-6 repeat the analysis of columns 2-4 but with an additional control for *Initial_Rating*. Note that there were no downward adjustments for supervisors giving low mean ratings. See the Appendix for variable descriptions.

[a] Omitted because of collinearity with additional, *Initial_Rating* variable.

Table 6
Survey data on evaluation system fairness, satisfaction, and favoritism

Panel A: Employee responses to survey questions

Junior, Staff, and Senior Auditor **questions:**	**Mean**	**SD**
a. The outcome of my most recent appraisal was fair. (*Outcome_Fairness*)	4.58***	1.85
b. The process that was used to determine the outcome of my most recent appraisal was fair. (*Process_Fairness*)	3.96	1.81
c. Overall, I am satisfied with (the performance evaluation system). (*Satisf_System*)	3.02***	1.85
d. Favoritism (not merit) gets people ahead. (*Favoritism*)	4.52***	1.84
Final_Shares of Survey Respondents	2.52	0.98

Supervisor **and** ***Program Director*** **Questions:**	**Mean**	**SD**
a. The outcome of auditors' most recent appraisal was fair. (*Outcome_Fairness*)	5.10**	1.62
b. The process that was used to determine the outcome of auditor appraisals was fair. (*Process_Fairness*)	4.67**	1.91
c. Overall, I am satisfied with (the performance evaluation system). (*Satisf_System*)	3.35**	1.97
d. Favoritism (not merit) gets people ahead. (*Favoritism*)	3.22**	2.16
Final_Shares of Survey Respondents	2.60	0.53

*** p<0.01, ** p<0.05, * p<0.10 indicates significance of the difference from the neutral response of 4 (two-tailed).

Panel B: Correlations between Employee Responses and their Performance Outcome

Auditors (n=221)

		1	2	3	4	5	6
1	*Final_Shares*						
2	*Outcome_Fairness*	**0.48**					
3	*Process_Fairness*	**0.32**	**0.75**				
4	*Satisf_System*	**0.32**	**0.49**	**0.62**			
5	*Favoritism*	**-0.25**	**-0.33**	**-0.39**	**-0.44**		
6	*Down_Adj*	**0.09**	0.09	0.06	0.08	-0.03	
7	*Up_Adj*	**0.09**	0.03	0.06	0.02	0.04	-0.05

Supervisors and Program Directors (n=48)

		1	2	3	4
1	*Final_Shares*				
2	*Outcome_Fairness*	**0.30**			
3	*Process_Fairness*	**0.34**	**0.80**		
4	*Satisf_System*	**0.31**	**0.58**	**0.57**	
5	*Favoritism*	**-0.29**	**-0.42**	**-0.55**	**-0.42**

Bolded Pearson correlations are significant at the 10% level (two–tailed).

Final_Shares = the number of shares assigned to each employee based on their final evaluation