July 2020

Narrative Feedback in Subjective Performance Evaluations: Do Ratings Change the Narrative?

Kyle Stubbs
University of Massachusetts Amherst

Narrative Feedback in Subjective Performance Evaluations: Do Ratings Change the Narrative?

A Dissertation Presented

By

KYLE STUBBS

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2020

Management

Narratives in Subjective Performance Evaluations: Do Ratings Change the Narrative?

A Dissertation Presented

By

KYLE STUBBS

Approved as to form and content by:

_____

M. David Piercey, Co-Chair

_____

Jeremiah Wayne Bentley, Co-Chair

_____

Yoon Ju Kang, Member

_____

Andrew Lind Cohen, Outside Member

_____

George R. Milne, Ph.D. Program Director
Isenberg School of Management

DEDICATION

To my wife and kids.

ACKNOWLEDGMENTS

I would like to thank my dissertation committee, M. David Piercey, Jeremy Bentley, Yoon Ju Kang, and Andrew L. Cohen, for their guidance with this project. They have each provided important comments and suggestions throughout this project. Also, each has been an excellent example of simultaneously being a good scholar and a good person.

I would like to thank my fellow accounting Ph.D. students that were willing to talk with me about this project all along the way. Your encouragement was vital. I would like to also thank the accounting department at University of Massachusetts Amherst for their feedback in multiple workshops as well as for providing funding for this research.

My family also deserves extensive praise for believing in me and being patient with me through this journey.

ABSTRACT

NARRATIVE FEEDBACK IN SUBJECTIVE PERFORAMNCE EVALUATIONS: DO RATINGS CHANGE THE NARRATIVE?

MAY 2020

KYLE STUBBS, B.S. BRIGHAM YOUNG UNIVERSITY – PROVO

MAcc, BRIGHAM YOUNG UNIVERSITY – PROVO

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: M. David Piercey and Jeremiah Wayne Bentley

Several high-profile companies are leading the charge to remove subjective performance ratings from their performance management processes leaving only narrative evaluations. Using two experiments, I investigate the effects of the ratings on narrative evaluations supervisors provide. In chapter 1, I test theory on supervisor goal attainment to learn how simultaneously providing a performance rating affects the narrative evaluation supervisors provide to employees. In supervisors' seeking of honesty and social cost reduction goals, I predicted the favorability of narrative evaluations to depend on the presence of ratings and the purpose of the performance evaluation. I used psychological licensing and process accountability theories to develop different expectations for the effects of ratings on narrative evaluations when the evaluation is used for coaching or for bonus purposes. Although I find minor evidence that narrative evaluations are more lenient when social costs are present, I fail to find support for the expected interactive effect of ratings and the purpose of the evaluation on leniency in narrative evaluations. I also provide some insights as to why these hypotheses were not supported. In chapter 2, I test theory on motivated reasoning to learn how simultaneously providing a performance rating affects the narrative evaluation supervisors provide to employees in the presence of directional goals. I predict and find that the favorability of narrative evaluations is influenced by directional goals, but the effect of those directional goals on narrative evaluation favorability is reduced when supervisors also provide a numerical rating of the employee's performance. The reduction

provides evidence that numerical ratings are less affected by directional goals than are narrative evaluations because of the precision and reduced ambiguity of meaning in number scales. This numerical rating then acts as a reasonableness constraint on free-form narrative evaluations which reduces the effect of directional goals on the favorability of the narrative. This study builds on the subjective performance evaluation literature by investigating the relationship between two forms of subjective evaluations: ratings and narratives. Also, this study provides important cautionary information to firms that have removed or are considering removing subjective performance ratings from their performance management systems.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

This dissertation contains the motivation, relevant theory, and documentation for two experiments seeking to understand the effects of numerical ratings on narrative evaluations in a performance evaluation setting. Chapter 1 documents the results of the experiment I defended as my dissertation proposal. However, as often happens in scientific inquiry, the results did not conform to my expectations. Unfortunately, the general lack of statistically significant results in the Chapter 1 experiment makes it difficult to understand whether the lack of results are evidence that the theory applied is not relevant to the setting or that the lack of results are evidence of experimental design issues. In a quest to understand the effect of numerical ratings on narrative evaluations, I took what I learned from the Chapter 1 experiment and designed a new study that investigates a variant of this primary research question with an experiment that addresses some of the methodological issues in Chapter 1. I document the results from this new experiment as Chapter 2.

The Chapter 2 experiment resulted in evidence supporting theory that numerical ratings do affect narrative evaluations in performance evaluations. I wrote Chapter 1 and Chapter 2 of this dissertation as independent manuscripts that can be read separately to allow the reader maximum freedom to read Chapter 1 or Chapter 2 independently, if so desired. However, if the reader is interested in a story of my lessons learned, reading Chapter 1 followed by Chapter 2 will demonstrate how issues identified in Chapter 1 were addressed in the experiment documented in Chapter 2. Readers will notice similarities in the motivation and theory used in Chapter 1 and Chapter 2 given the overlapping nature of the content. Overall, I hope this dissertation can be useful both as a set of independent manuscripts as well as a linear story of how I sought to answer my research question.

CHAPTER 1

THE EFFECTS OF RATINGS AND EVALUATION PURPOSE ON NARRATIVE EVALUATIONS

1.1 Introduction

In the recent upheaval of traditional performance appraisal systems, some high-profile companies have removed numerical performance ratings from their appraisal process (Rock and Jones 2015; Silverman 2016). While the effectiveness of this shift is questioned by some proponents of ratings due to beliefs that ratings are useful despite having faults (Adler et al. 2016), we don't know how ratings (or the lack of ratings) impact the ever-present narrative evaluations provided to employees as part of the management control system. This practical dilemma facing companies raises an interesting question. How does quantifying a subjective performance evaluation with a rating affect detailed subjective narrative evaluations that traditionally accompany the ratings? Previous research on performance evaluation focuses almost exclusively on the performance rating as the outcome of interest, leaving the narrative evaluation under researched (Wilson 2010; Speer 2018). I perform a controlled experiment to fill this gap in practical and theoretical knowledge regarding potentially unintended effects of quantifying a subjective performance evaluation in a rating.

Subjective performance evaluations provide a way for firms to capture, store, and use information about employee performance that is difficult to objectively measure. Subjectively interpreting raw performance information into summarized evaluations can play a feedback role and a compensation-contracting role in firms' management control systems. Although numerical subjective performance measures play a significant role in determining important outcomes like the administration of compensation contracts (Bol 2008), these ratings are not without issues. Leniency (i.e., overly favorable evaluations) when evaluating poor performers is one of the most commonly documented problems with performance evaluations (e.g., Jawahar and Williams 1997; Moers 2005; Bol 2011; Bol, Kramer, and Maas 2016). Lenient evaluations provide

inaccurate information to both employees and the control system which can compromise decision-making in the firm. Previous accounting studies have documented changes in leniency in performance ratings due to changes in the control system environment (Moers 2005; Bol 2011). While most of the literature on performance appraisal focuses exclusively on numerical ratings[1], understanding leniency in narrative evaluations is important because these narrative portions of evaluations are known to be important to employees and can provide detailed information that employees can use to improve performance (Smither and Walker 2004). Also, for firms that had removed performance ratings, narrative evaluations are left alone as the formal evaluation method for employee feedback and provide the basis for subjective compensation contracting.

The debate over getting rid of performance ratings makes it clear that we need to understand the information that is transmitted to employees and decision-makers outside of the numerical rating (Adler et al. 2016). I use a goal attainment perspective to formulate expectations of evaluator behavior during subjective performance evaluations as recommended by Murphy and Cleveland (1995). Supervisors have at least two competing goals when completing performance evaluations: honesty and social cost reduction. Leniency is evidence of supervisors' desire to reduce social costs associated with giving accurate, low performance evaluations (Jawahar and Williams 1997). I expect the ability to and strategies used to satisfy these two competing goals will be impacted by the mix of numerical and narrative portions in the performance evaluation and the purpose of the evaluation. Consistent with prior research, I expect narrative evaluations without ratings to be more lenient as the social costs related to the evaluation increase. While any social costs should increase leniency to at least partially satisfy the social cost reduction goal, the level of social costs should change depending on how that evaluation will be used. I investigate

[1]Hereafter, numerical ratings will be labeled ratings. While the ratings specifically investigated in this study contain numbers, ratings can take many forms including those with and without numbers (Murphy and Cleveland 1995). A numberless rating could be as simple as a 3-category choice of needs improvement, meets expectations, or exceeds expectations. These types of ratings share many important attributes with numbered ratings including comparability and forced categorization and would be expected to result in similar effects as those theorized in this study.

evaluations used in their two primary purposes: compensation contracting and employee feedback. Consistent with prior literature, I expect the greatest social costs and most lenient evaluations to be for those used in compensation contracting because of the additional social cost of an employee's compensation being affected by the evaluation (Jawahar and Williams 1997).

When considering the impact of ratings on narrative evaluations, I separately develop expectations of the effect of ratings on narrative evaluations under the two different purposes since prior research suggests that rater strategies change under different purposes (Zedeck and Cascio 1982). For feedback purposes (e.g., used to coach an employee), the addition of a rating provides an additional communication method for supervisors to use to satisfy their competing goals of honesty and social cost reduction. Theory on psychological licensing suggests that supervisors may seek to obtain a license (i.e., reduce social costs) to better achieve the honesty goal of giving an honest, unfavorable evaluation of a poor performer (Miller and Effron 2010; Merritt et al. 2012). Due to the relative strengths and weaknesses of ratings and narrative evaluations (Rock, Davis, and Jones 2014), supervisors will use the rating to provide a favorable picture of employee performance, thereby obtaining a license to be critical and honest in the narrative evaluation. In summary, for evaluations with a feedback purpose only, this theory predicts the addition of a rating will reduce the leniency of narrative evaluations.

For purposes of compensation contracting (e.g., used to determine an employee bonus), the fact that the evaluation will be provided to a downstream decision-maker to determine the compensation outcome for the employee alters the way adding a rating will affect narrative evaluations. The involvement of the downstream decision-maker will increase a form of process accountability for the supervisor's evaluation (David 2013; Erdogan 2002; Lerner and Tetlock 1999). Specifically, this process accountability will lead them to match the favorability of their narrative evaluations with their rating to give the impression that the evaluation was done appropriately. While the rating and written evaluation will match, the addition of a rating increases the social costs of giving honest feedback because of the clear, unambiguous link

between the rating and the employee's valued outcome (e.g., bonus). Narrative evaluations on the other hand require substantial interpretation and are difficult to compare across individuals making the link between the narrative evaluation and the employee's valued outcome ambiguous (Speer 2018; Adler et al. 2016). With the increased social costs of giving honest ratings that clearly link to valued outcomes for the employee, narrative evaluations accompanied by a rating should be more lenient than narrative evaluations without a rating.

To address my research question and test these expectations, I administer a 2x3 between-subjects experiment manipulating rating type and evaluation purpose where graduate business student participants take the role of a supervisor at a hypothetical company and evaluate an employee's performance. After each participant views a video of a moderately poor performer, they must fill out a performance evaluation for the employee. Since I am focused on understanding how narrative evaluations change under different conditions, all these conditions require participants to complete a narrative evaluation for the employee. The rating type is manipulated such that the evaluation either requires no rating or does require participants to rate employee performance on a 101-point scale. I manipulate performance evaluation purpose at three levels: compensation contracting (i.e., evaluation affects employee bonus likelihood), feedback only (i.e., evaluation used only for coaching the employee), and no purpose (i.e., evaluation is private, meaning it is not used in the performance management system nor shown to the employee). The dependent variable of interest is the *Narrative Evaluation Favorability* measured multiple ways using a manually coded global judgment of favorability, sentiment analysis using machine learning derived sentiment scores (IBM's Watson), and the manually coded number of positive statements less the number of negative statements in each evaluation.

In my experiment, I expected to find public narrative evaluations (without ratings) are more lenient than private narrative evaluations with narrative evaluations used to determine a bonus (i.e., compensation contracting) being the most lenient. While I do find that public narrative evaluations (without ratings) tend to be more lenient than the control conditions, I do

not find that narrative evaluations are the most lenient when used to determine a bonus. With regards to the effect of ratings on leniency in narrative evaluations, I find no evidence of the predicted interaction between the presence of a rating and the purpose of the evaluation. The results also suggest that participants did not attend to nor understand the manipulation of evaluation purpose. This may be an explanation for the lack of results in this study.

This study builds theory to contribute to the heated debate on the removal of performance ratings (Adler et al. 2016; CEB 2016; Cappelli and Tavis 2016) by investigating one potential impact of this trend. I expected to show that the presence of ratings has an impact on the accuracy of narrative evaluations. However, I fail to find to evidence of this effect. Based on my expected results, I had hoped to provide evidence that contributes to practice in multiple ways. First, I expected to find that narrative evaluations with employee incentives tied to them would be more accurate without a rating compared to those with a rating. For firms concerned about development of employees through feedback, I expected to find that narrative evaluations used just for coaching may be less accurate (more lenient) without a performance rating. Although this study did not provide evidence of this theory, the question of how ratings impact narrative evaluations remains an important and open question that should be investigated.

I also attempted to contribute to the research on subjective performance evaluation by deepening the collective understanding of factors that affect narrative evaluations. Subjective narrative evaluations have received very little attention in the performance evaluation literature (Wilson 2010; Speer 2018). This study describes an experiment with expectations that narrative evaluations are not simply qualitative representations of their quantitative counterparts. However, I fail to find evidence that narratives and quantitative evaluations are unique and are affected by evaluator goals differently.

I also extend the management accounting literature on performance feedback by looking into determinants of performance feedback rather than outcomes of feedback. While substantial literature has focused on the effects of objective control system feedback on employee effort

(e.g., Hannan, Krishnan, and Newman 2008; Hannan, McPhee, Newman, and Tafkov 2013; Casas-Arce, Lourenço, and Asís Martínez-Jerez 2017), very little has investigated how managers determine what performance feedback to give.

Last, I contribute to the broader accounting literature on narrative communication (e.g., Li 2008; Asay, Libby, and Rennekamp 2018; Bentley 2019). These studies investigate how agents communicate with principals using narrative reports and how it affects the agents and principals. I extend the literature by investigating how principals communicate using narrative reports to agents. Specifically, I investigate how principals' narrative evaluations of agents are affected by quantitative ratings. However, I do not find support for this predicted effect.

This paper continues with Section II establishing the background and developing the hypotheses. Section III and Section IV describe the methodology used to investigate the research question and the results, respectively. Section V concludes the paper.

1.2 Background and Hypotheses

1.2.1 Background on Subjective Performance Evaluation

When performance is not easily objectively measurable, using observers of performance to subjectively evaluate the performance of employees provides essential information to the management control system. This subjective performance evaluation fills in gaps left by objective measures to achieve multiple roles of control systems (Demski and Feltham 1976). Subjective performance evaluation plays a role in compensation contracting through its use in performance-related decisions like promotion, bonus allocation, and raise determination (Bol 2008). Research in management accounting has provided important insights into the use of subjectivity in these types of compensation contracting decisions (e.g., Bol 2011; Bol and Smith 2011; Chen, Jermias, and Panggabean 2016). Subjective performance evaluation also fills a feedback role in control systems by providing crucial information on employee performance for the employee to develop and better understand how their performance compares to firm expectations (London 2003). The feedback literature in management accounting largely focuses on how different characteristics of

objective feedback not influenced by supervisor subjectivity impact employee performance. Recent studies investigating this objective feedback include those on relative performance information (Hannan et al. 2008; Hannan et al. 2013) and absolute performance information (Casas-Arce et al. 2017). While these studies emphasize the importance of feedback in management accounting, they focus on information that is not subjectively generated by supervisors. Performance that is difficult to measure objectively will still need to be subjectively evaluated by supervisors to give employees the feedback they need. Thus, understanding determinants of subjective feedback is important because of its effects on employees.

While the past literature on subjective performance evaluation has focused almost exclusively on numerical performance information, narrative or verbal evaluations play an important role in control systems. In traditional performance appraisal used for compensation contracting, supervisors provide both performance ratings and narrative evaluations of employee performance (Gorman, Meriac, Roch, Ray, and Gamble 2017). The content of these narrative summaries remains a new and under-researched area of performance evaluation (Speer 2018; Wilson 2010) even though employees report paying much attention to the narrative feedback (Smither and Walker 2004). Narrative performance evaluations allow for richness and nuance in communication unavailable using performance ratings that can be useful to employees and decisionmakers. Given the importance of narrative feedback in practice and the recent removals of performance ratings mentioned in the introduction of this paper, understanding how narrative evaluations change with and without performance ratings used for either compensation contracting or feedback purposes is a timely and important question.

1.2.2 Supervisor Goal Attainment in Performance Evaluation

To understand how performance ratings might impact narrative performance evaluations, I rely on the goal attainment framework of subjective performance evaluation presented by Murphy and Cleveland (1995). One of the important tenets of this framework is that supervisors have goals besides being perfectly accurate in their evaluations. Given the social dynamic present

in supervisor-subordinate relationships, supervisors also have goals to reduce social costs associated with giving honest but unfavorable evaluations to subordinates. Evidence of this goal to reduce social costs is apparent in one of the common behavioral tendencies of evaluators in subjective performance evaluation, leniency (Jawahar and Williams 1997; Moers 2005; Bol 2011; Bol et al. 2016). Leniency is the tendency for an evaluator to give an employee an evaluation reflecting better performance than would be expected given actual employee performance (Decotiis and Petit 1978). Managers are lenient in evaluations to avoid damaging relationships, having awkward or uncomfortable conversations, hurting employee's promotion or bonus prospects, and experiencing retaliatory behavior from employees (Murphy and Cleveland 1995). These social costs are particularly potent for employees that are underperforming because honest evaluations for these employees are more likely to bring on these social costs. Giving honest evaluations to high performers is unlikely to incur these social costs since the honest evaluation would be received favorably by the employee. Since this paper investigates variations in leniency, I focus this theoretical development on the process of evaluating an employee that has underperformed.

Lenient performance information is a concern for both compensation contracting and employee feedback in the firm. For compensation contracting purposes, lenient evaluations of poor performers can lead to evaluation compression which makes it more difficult to fairly distribute contracted rewards and make other personnel decisions like promotion which can impact employee motivation negatively (Bol et al. 2016). For feedback purposes, lenient evaluations reduce opportunities for employees to develop and learn. Since performance feedback information has long since been thought to be useful for performance improvement (Kluger and DiNisi 1996), poor quality performance information has the potential to restrict employee development and success.

1.2.3 Social Costs of Narrative Evaluations

While the literature on leniency in evaluations of poor performers has focused on the leniency of numerical ratings, many of the aforementioned social costs of honest evaluations would be expected to also exist for narrative evaluations even in the absence of ratings. Rather than provide completely honest narrative evaluations, I expect supervisors to balance the competing goal of reducing social costs by being lenient in narrative evaluations. Murphy and Cleveland (1995) suggest that inaccurate ratings are not necessarily evidence of an inability to accurately judge performance. Instead, it is evidence that managers have goals other than being perfectly accurate in evaluations. Thus, a private evaluation (i.e., visible only to the evaluator) devoid of social costs could be viewed as a benchmark of honesty and accuracy. Once social costs are applied to the setting by making it public and for use in the control system, I expect narrative evaluations to be lenient.

To the extent that managers have a goal to reduce social costs, leniency should increase with social costs. These social costs are likely to depend on what the evaluation is used for in the control system. Prior research on performance ratings finds that ratings used for compensation contracting (administration of organizational rewards) are more lenient than those used solely for feedback purposes (development of employees) due to the greater social costs of giving honest evaluations under compensation contracting (Jawahar and Williams 1997). Specifically, costs associated with harming or failing to help someone achieve a valued outcome, like a bonus, become important considerations when the evaluation is used for compensation contracting. I expect these higher social costs associated with compensation contracting to exist when evaluations consist of a narrative which replicates prior literature but with narrative evaluations rather than ratings.

In summary, relative to a private narrative evaluation benchmark, I expect a narrative evaluation used for either compensation contracting or feedback purposes (i.e., a public narrative evaluation) to be more lenient due to the presence of social costs. Also, I expect narrative

evaluations used for compensation contracting to be more lenient than narrative evaluations only used for feedback purposes. This hypothesis establishes the effects of social costs on narrative evaluations without ratings.

H1a: Without ratings, supervisors will provide more lenient narrative evaluations of employee performance when the evaluation is either used in compensation contracting or coaching rather than when it is kept private (no purpose).

H1b: Without ratings, supervisors will provide more lenient narrative evaluations of employee performance when the evaluation is used for compensation contracting rather than used for feedback only.

While H1a and H1b are straightforward predictions from existing performance evaluation literature, they help establish benchmarks from which to see how a numerical rating affects leniency. In the following subsections, I develop theory to explore how a numerical rating might affect the leniency of narrative evaluations. However, as described in the development of H1b, social costs and motivations for performance evaluation are different under different purposes. Literature on performance evaluation purposes from outside of accounting suggest that rater strategies are different when evaluations are used for different purposes (Zedeck and Cascio 1982; Jawahar and Williams 1997). Given these differences, I develop expectations for the effect of ratings on narrative evaluation separately by purpose.

1.2.4 Effect of Ratings on Narrative Evaluations – Employee Feedback Purpose

Since supervisors have competing goals of honesty and social cost reduction, an optimal performance evaluation would simultaneously and fully satisfy both goals. However, up to this point, I have established that at least some compromise or an attempt to satisfy the social cost reduction goal occurs resulting in lenient narrative evaluations. When I consider the role a rating might play in satisfying these goals, I present the possibility that the additional communication method (i.e., the rating) may allow for supervisors to satisfy both competing goals simultaneously

using the two communication methods. If this could occur, supervisors would maximize goal satisfaction by doing so.

Theory on psychological licensing provides a basis for why a supervisor may be able to better satisfy both goals simultaneously when an additional communication method is available. A psychological license is a person's "perception that they are permitted to take an action or express a thought without fear of discrediting themselves" (Miller and Effron 2010, p116). While moral licensing is a common form of psychological licensing, behavior need not be morally discrediting to require a license (Miller and Effron 2010). Psychological licenses remove barriers to taking a desired action. Since supervisors are concerned with how an employee will receive honest, unfavorable feedback, a psychological license to be honest and critical would remove the barrier and result in a greater ability to achieve the honesty goal. Research on psychological licensing through group membership provides some evidence that supervisors may be able to earn a license to be critical by being lenient in part of their evaluation. Specifically, relative to non-group members, people identified as group members are licensed to be more critical of their own group members (Hornsey, Oppes, and Svensson 2002; Hornsey and Imani 2004). These people obtained the psychological license to be critical through their group membership and reactions to their criticism were more favorable due to greater perceived constructiveness of feedback. Similarly, I expect a supervisor to be able to earn a license to be critical and honest with the employee by establishing themselves as unbiased and "on the side" of the employee. I propose that they can achieve this by providing a salient favorable evaluation to the employee using one of the pieces of the evaluation. Doing so would allow the supervisor to give an unrestricted (psychologically licensed) honest evaluation using the other piece of the evaluation. While most of the early literature on psychological licensing focuses on behavior after people already feel licensed, research also finds that people proactively seek out licenses to remove barriers associated with potentially discrediting behavior (Merritt et al. 2012). Thus, I expect that

supervisors will seek out this license to be honest and it would be perceived as an effective license.

I expect managers to select which of the pieces of the evaluation, the rating or the narrative evaluation, to use to satisfy each goal by which piece is better suited for each task. Ratings provide summarized global judgments of employee performance and create strong emotional responses (Rock et al. 2014), while narrative evaluations provide the opportunity for detailed feedback that the employee can use to improve. Ratings' salience and impact on emotions allow it to be an effective license to establish the supervisor as being on the employee's side. The narrative evaluation naturally provides the most value in providing honest feedback since it can provide actionable information for the employee to use to improve performance which is the purpose of feedback and one of the motivators for the honesty goal. In summary, I expect supervisors to use lenient ratings to obtain a license to be honest and critical in their narrative evaluations. With this license, I expect narrative evaluations accompanying a rating to be more honest (less lenient) than narrative evaluations with no rating.

H2a: When evaluations will be used only for feedback purposes, narrative evaluations accompanied by a rating will be less lenient than narrative evaluations with no rating.

1.2.5 Effect of Ratings on Narrative Evaluations – Compensation Contracting Purpose

Subjective performance evaluations for a compensation contracting purpose have the additional social cost of affecting valued outcomes for the employee (Jawahar and Williams 1997). For the evaluation to be used to determine this valued outcome it must become part of the control system and be used by a compensation decision-maker to decide on valued outcomes like allocating a bonus. I expect this information transfer to a compensation decision-maker to impact the way supervisors prepare their evaluations that include ratings. First, by transferring the information to a decision-maker, supervisors will be accountable to the decision-maker to do a good job or at least appear to have done a good job with their evaluation. The type of

accountability created by oversight depends on the context and information available to the decision-maker (Lerner and Tetlock 1999). Research on accountability bifurcates accountability into two types relevant for this study: outcome accountability and process accountability (Siegel-Jacobs and Yates 1996; Lerner and Tetlock 1999). In the creation of subjective performance evaluations, outcome accountability (i.e., responsibility for outcome accuracy) is difficult to enforce since the decision-maker is unaware what an accurate evaluation outcome would be because he or she did not personally observe the performance. However, the decision-maker can attempt to assess process accountability (i.e., responsibility for process used to make a decision) by investigating whether proper decision strategies have been taken by the supervisor in making the evaluation. Since the decision-makers lack underlying performance information, they can only assess the appropriateness of the evaluation using the rating and narrative portion of the evaluation provided by the supervisor. Evidence from performance evaluation research suggests that narrative evaluations can play the role of satisfying process accountability by justifying or explaining the rating (David 2013; Erdogan 2002). Thus, different from evaluations used only for feedback purposes, supervisors will need to match the favorability of ratings and narrative evaluations to create perceptions that the evaluation process was appropriate.

Along with encouraging matching between ratings and narrative evaluations, the additional layer of communication to the compensation decision-maker will affect the social costs associated with the type of evaluation provided. Ratings and narrative evaluations have fundamental differences that affect their suitability for use in distributing compensation rewards. Researchers and practitioners recognize the inherent difficulties with using narrative evaluations as the sole basis of bonuses or promotion (Adler et al. 2016). Quantitative ratings are easier to compare between individuals than narrative evaluations which makes them the natural choice for administering reward contracts (Speer 2018). Narrative evaluations are free-form and require significant interpretation to be used for administering contracts or determining promotions. Therefore, the causal relationship between the favorability of a narrative evaluation and the

employee's valued outcome is much more ambiguous than the causal relationship between a rating and the valued outcome. Since supervisors are concerned with incurring social costs associated with impacting employee's valued rewards, I expect the social costs of being honest to be higher for ratings than for narrative evaluations due to the clear relationship between ratings and the administration of the valued outcome. In summary, a supervisor providing a rating will be pressured to be more lenient in the evaluation because of how clearly it will affect the employee's compensation relative to a narrative evaluation by itself. Also, consistent with process accountability pressure, I expect the narrative evaluation to match the lenient rating. In summary, when evaluations are used in compensation contracting, I expect supervisors to provide more lenient narrative evaluations in the presence of a rating compared to when no rating is present.

H2b: When evaluations will be used for compensation contracting purposes, narrative evaluations accompanied by a rating will be more lenient relative to narrative evaluations with no rating.

1.3 Methodology

To test my hypotheses, I perform an experiment where participants act as supervisors with the role of subjectively evaluating the performance of an employee.[2] Using a 2X3 between-subjects design, I manipulate *Rating Type* at two levels (No Rating or Rating) and *Evaluation Purpose* at three levels (Compensation Contracting purpose, Feedback purpose, No purpose).[3] I administer this experiment to 144 graduate business students at a large public university in the northeastern United States as participants[4]. Participants were on average (median) 35.5 years old

[2] This study was approved by the Institutional Review Board at the author's institution.
[3] I collect data for three additional conditions (72 additional participants) for use in supplemental analyses to provide additional evidence of the theory proposed in this paper. Each of these conditions focus on ratings only and do not include a narrative evaluation. The three conditions are as follows: 1) Rating with Compensation Contracting Purpose, 2) Rating with Feedback Purpose, and 3) Rating with No Stated Purpose. These conditions provide benchmarks to investigate how ratings are changed when accompanied by narrative evaluations.
[4] Most participants (95.14%) received course credit for participation in the study. A small percentage (4.86%) of participants did not receive extra credit for participation, instead they voluntarily participated without compensation.

(35 years old) and had an average (median) of 43.9 (24) months of supervisor experience. 79.2% of participants reported listening to music at least daily. With the combination of supervisor experience and high exposure to music, I believe these graduate business students are appropriately knowledgeable to participate in this study (Libby, Bloomfield, and Nelson 2002).

1.3.1 Procedures

After accessing the study using a Qualtrics survey link, participants take the role of a supervisor in a hypothetical cruise ship vacation company. After reading basic information about the company, the participant learns they need to evaluate an employee's performance from a recent cruise. The employee, Spencer, sings and plays music to provide entertainment for cruise guests. After being introduced to Spencer through a brief video, participants are instructed about the form the evaluation will take. Participants then view a 3.5-minute video of Spencer's performance and can take notes on Spencer's performance for use later when they fill out the evaluation form. After viewing the performance, they evaluate Spencer using the performance evaluation form based on their condition. Finally, participants answer additional questions to help me better understand why they behaved the way they did and to collect demographic information.

1.3.2 Task Performance Video

I used a video of a musical performance as the performance for supervisors to evaluate for specific experimental design purposes. First, a musical performance is an inherently subjective task. Since subjectivity is at the center of subjective performance evaluation and is essential to the theory presented in this study, a highly subjective performance context like a musical performance is warranted. Second, real world performance evaluations are based on a rich set of information which allows for significant variation in narrative evaluations. Experimental instruments traditionally used for investigating variation in performance ratings are not sufficiently rich to allow for adequate variation in narrative evaluations. To balance the time constraints of participants and the richness of data to evaluate, I chose a task that is both brief and rich with data to evaluate. By viewing the performance of a single song, evaluators can focus on a

variety of aspects: audience interaction, vocal quality, instrumentation quality, song choice, mistakes, stage presence, etc. This task provides an effective and efficient means of testing the hypotheses. Last, music is a task that many people are familiar with. It can be difficult to find a sufficient pool of participants that are able to write intelligently about and evaluate performance on a specific task. Given the ubiquity of music, I was able to efficiently access participants that were reasonably equipped to write about an employee's musical performance.

To construct a musical performance that was appropriate for this study, I hired an actor/musician to be the employee, Spencer. Since my theoretical expectations require social costs of honest evaluations to be a concern to supervisors, the musical performance needs to be moderately poor for a professional musician. I hired a professional musician to perform a popular song with seeded mistakes in the performance (e.g., missed lyrics, piano mistakes, failure to interact with the audience). This professional musician was compensated for her acting performance and was also eligible for real bonus compensation based on the evaluations of the participants and the decision-making of the boss (see discussion of the *Evaluation Purpose* manipulation for more details).

1.3.3 Manipulations

1.3.3.1 Rating Type

Rating type is manipulated at two levels. Participants in the no rating condition, *No Rating*, are not required to provide any numerical rating in their performance evaluation. Participants in the rating present condition, *Rating*, are asked, "Please rate the performance of your employee, Spencer, on a scale from 0 to 100 based on your expectations for a professional performer at Riverboat Cruises Inc."[5] Participants respond on a sliding scale with a number from 0 to 100 with labels at 0 ("Far Below Expectations"), 50 ("Meets Expectations"), and 100 ("Far

[5] While performance ratings can take the form of absolute or relative formats (like rankings), absolute ratings appear to be more commonly used in practice (Gorman et al. 2017). I use an absolute rating scale in this study for external validity and simplicity with supervisors only needing to evaluate a single employee.

Above Expectations"). In the *Rating* conditions, the performance rating scale is presented simultaneously with the narrative evaluation. The rating is at the top of the page and the narrative evaluation follows.

1.3.3.2 Evaluation Purpose

Evaluation purpose is manipulated at three levels (Compensation Contracting purpose - *Bonus*, Feedback purpose - *Coaching*, No purpose – *No Purpose*). The compensation contracting purpose is operationalized by tying the overall performance evaluation to a bonus for the employee. In this *Bonus* condition, supervisors learn that the employee, Spencer, is eligible for a $50 performance bonus and that their evaluation will affect the likelihood of Spencer receiving the $50 bonus. Specifically, they are told the following, "This performance evaluation will be used in making promotion, bonus, and job security decisions, in addition to coaching Spencer. Specifically, Spencer is eligible for a $50 performance bonus. Your evaluation will be sent to your boss who will determine whether Spencer receives the bonus." Participants in these conditions are also told, "Your evaluation will be sent to Spencer and your boss." Although the cruise company is hypothetical, participants were told that the employee, the boss, and the bonus were real. After the experiment was complete, I sent all evaluations from the *Bonus* conditions to a real person[6], the boss, who used the evaluations to decide whether Spencer, the singer in the video, received the bonus. After receiving all the evaluations from the *Bonus* conditions, the boss in this study decided not to award the employee, Spencer, the $50 bonus.

The feedback purpose is operationalized by describing the purpose of the overall performance evaluation as just for coaching the employee. In this *Coaching* condition, supervisors are told the following, "This performance evaluation is just for the benefit of

[6] The boss in this study was a PhD student unaffiliated with this study. The boss did not personally know the employee, Spencer, nor did the real person acting as the boss have to pay the $50 bonus if the decision to give the bonus was made. The boss was simply asked to read the evaluations and decide whether Spencer should be awarded the bonus. There is no deception in this study.

coaching Spencer and will not be used in making promotion, bonus, and job security decisions for Spencer". Participants are also told, "Your evaluation will only be sent to Spencer."[7]

The *No Purpose* conditions require participants to provide a private evaluation of the employee. Participants are instructed the following way, "Your evaluation will not be sent to Spencer or anyone else at Riverboat Cruises Inc. Your evaluation will be kept private." This condition is intended to capture a baseline for honesty in narrative evaluations by removing the presence of social costs. The performance evaluation has no indicated purpose since it will not be communicated to anyone.

1.3.4 Dependent Variables

For narrative evaluations, participants responded to the following prompt, "Please provide a written evaluation of your employee, Spencer, based on your expectations for a professional performer at Riverboat Cruises Inc [Please write a minimum of 250 characters]." Participants were forced to write at least 250 characters to ensure that the text responses would be long enough to allow for measurable variation in the construct of interest. I use the narrative evaluations as the raw material to create the primary dependent variable.

Since I am interested in understanding how leniency of narrative evaluations changes under different conditions, I focus on the construct of *Narrative Evaluation Favorability*. I measure this construct in multiple ways. For use in the primary analyses, I code each narrative evaluation while blind to experimental conditions on the same 101-point scale used in the *Rating* manipulation (0-Far Below Expectation, 50-Meets Expectations, 100-Far Above Expectations). This captures a holistic measure of the favorability of the narrative evaluation absent any rating information. In supplemental analyses, I use additional measures of *Narrative Evaluation Favorability*. As one alternative measure of the *Narrative Evaluation Favorability*, I use the

[7] To increase the felt social costs in this study, I mix both hypothetical and real elements into the experiment. While the cruise ship scenario is hypothetical, the impact of the supervisor's evaluations on the employee is real. In all conditions other than the no stated purpose conditions, the evaluations were provided to the employee, Spencer, by email after the study was over. There is no deception in this study.

natural language understanding service from IBM Watson[8]. This service uses deep learning to analyze text and provide metadata including the sentiment score of the data. The sentiment score is bounded by -1 and 1 with higher numbers being interpreted as having a more positive sentiment. I use the sentiment score for each narrative evaluation as an alternative measure of *Narrative Evaluation Favorability*. As a third measure of *Narrative Evaluation Favorability*, I code the number of positive and negative statements in each narrative evaluation while blind to conditions. I then calculate the number of positive statements minus the number of negative statements for each narrative evaluation. This captures the overall favorability of each evaluation by counting the valence of the underlying contents of the evaluation. For all three measures of *Narrative Evaluation Favorability*, I interpret higher values as being indicative of more favorable evaluations.

1.4 Results

1.4.1 Manipulation Checks

Before proceeding with the hypothesis tests, I check whether participants understand the *Evaluation Purpose* manipulation. Participants responded to two questions to capture their understanding of the *Evaluation Purpose* manipulation. First, they respond to the following question with a "yes" or "no" answer: "Is your employee, Spencer, going to see the performance evaluation you provided?". Second, they respond to another "yes" or "no" question, "Is your boss at Riverboat Cruises going to see the performance evaluation you provided and use it to determine Spencer's bonus?" Participants in the no purpose conditions should answer "no" to both questions. Participants in the coaching condition should answer "yes" to the first question and "no" to the second question. Participants in the bonus condition should answer "yes" to both questions. Only 57.6% (83 out of 144) of participants correctly answered these manipulation check questions. The manipulation check failures primarily occurred in the no purpose and

[8] See https://www.ibm.com/cloud/watson-natural-language-understanding for more details.

coaching conditions with only 23.4% (11 out of 47) and 65.31% (32 out of 49) passing the manipulation check questions, respectively. 83.33% (40 out of 48) of participants in the bonus conditions accurately responded to both questions suggesting that participants in these bonus conditions generally understood their evaluation type. Overall, the low success rate is a cause for concern when generalizing the results from this study. The presence or lack of results in the following sections could be due to participant's lack of attention or understanding of the manipulations rather than evidence supporting or refuting the theory underlying the hypotheses. I present my analyses using the full sample of 144 participants.[9]

1.4.2 Hypothesis Tests

A graphical representation of my hypotheses is included as Figure 1. I first investigate my hypotheses using *Narrative Evaluation Favorability* as coded by me on a 0 to 100 scale with 0 labeled as "Far Below Expectations", 50 labeled as "Meets Expectations", and 100 labeled as "Far Above Expectations". I tabulate descriptive statistics of *Narrative Evaluation Favorability* by condition in Table 1, Panel A. I also graph the cell means in Figure 2. In Table 1, Panel B, I begin formally testing my hypotheses by documenting a two-way Analysis of Variance (ANOVA) of *Narrative Evaluation Favorability* with *Rating Type*, *Evaluation Purpose*, and the interaction as factors. As initial evidence of H1a and H1b, I find a significant effect of *Evaluation Purpose* on *Narrative Evaluation Favorability* ($F_{2,138}$=5.840, p<.01). However, I do not find evidence of the interaction implied in H2a and H2b ($F_{2,138}$=0.333, p=0.718).

I follow up the significant effect of *Evaluation Purpose* by performing the specific simple effect tests predicted in H1a and H1b. As predicted in H1a, *Narrative Evaluation Favorability* without a rating is higher in the coaching condition than the no purpose condition

[9] I replicate the primary analyses of the *Narrative Evaluation Favorability* variable with each of the three measures of the construct (Coded Favorability, IBM Watson Sentiment, and Positive Statements minus Negative Statements) after dropping manipulation check failures. In all three ANOVAs, neither independent variable nor the interaction between the two is statistically significant (all p's > 0.38). However, given the large decrease in statistical power due to the reduction in number of participants, the insignificant results could be due to a lack of statistical power.

(estimate=7.625, $t_{df=138}$=2.111, two-tailed p=0.037). Although directionally consistent with H1a, *Narrative Evaluation Favorability* without a rating is not significantly higher in the bonus condition than the no purpose condition (estimate=3.926, $t_{df=138}$=1.109, two-tailed p=0.270). Thus, I find mixed results for H1a. H1b is not supported as I find the bonus condition without a rating is directionally less favorable than the coaching condition without a rating although this effect is not significant (estimate=-3.699, $t_{df=138}$=1.044, two-tailed p=0.298). This evidence is in the opposite direction of the effect predicted in H1b. The findings show partial support for H1a and fail to support H1b. Since the interaction term in the ANOVA was insignificant and the visual fit of the predicted interaction is poor, I do not perform planned contrasts testing for H2a and H2b. H2a and H2b are not supported.

In summary, I find some evidence consistent with prior literature suggesting that introducing social costs leads to more lenient evaluations, in this case higher *Narrative Evaluation Favorability* (Jawahar and Williams 1997; Moers 2005). However, I fail to find evidence consistent with prior research that social costs are higher when compensation is impacted by an evaluation compared to when it is just used for employee feedback (Jawahar and Williams 1997). Also, I do not find evidence that ratings impact *Narrative Evaluation Favorability* and interact with *Evaluation Purpose*.

1.4.3 Supplemental Analyses

1.4.3.1 Alternative Measures of Narrative Evaluation Favorability

As an alternative measure of *Narrative Evaluation Favorability*, I use the sentiment score provided by IBM Watson's natural language understanding software for each narrative evaluation. I replicate the analyses in Table 1 using this alternative measure and tabulate the descriptive statistics, corresponding ANOVA, and simple effect tests in Table 2. I also graph the cell means in Figure 3. The results are similar to the findings for the primary measure of *Narrative Evaluation Favorability* in Table 1, although the results here are less statistically significant. Specifically, I find a marginally significant effect of *Evaluation Purpose* on *Narrative*

Evaluation Favorability ($F_{2,138}$=2.555, p=0.081). Also, similar to the primary analysis in Table 1, there is no interaction effect between *Evaluation Purpose* and *Rating Type* on *Narrative Evaluation Favorability* ($F_{2,138}$=0.038, p=0.963). Following up with simple effects tests for H1 in the absence of a rating, *Narrative Evaluation Favorability* marginally significantly increases for those in the coaching condition compared to the no purpose condition (estimate=0.189, $t_{df=138}$=1.413, one-tailed p=0.080), but *Narrative Evaluation Favorability* is not significantly higher in the bonus condition than the no purpose condition (estimate=0.084, $t_{df=138}$=0.639, two-tailed p=0.524). This provides partial evidence of H1a with participants in the coaching condition without a rating having higher *Narrative Evaluation Favorability* than those in the no purpose condition without a rating. H1b is not supported with the bonus condition without a rating being directionally lower but not statistically different from the coaching condition without a rating (estimate=-0.105, $t_{df=138}$=-0.802, two-tailed p=0.424).

In untabulated analyses, I analyze the *Narrative Evaluation Favorability* using the number of positive statements minus number of negative statements. Using an ANOVA, I find results similar to those documented in Table 1 and Table 2 with respect to my hypotheses. I find a marginally significant main effect of *Evaluation Purpose* on *Narrative Evaluation Favorability* ($F_{2,138}$=2.947, p=0.056). I find overall evidence that social costs influence *Narrative Evaluation Favorability* but fail to find evidence that the bonus condition is more positive than the coaching condition. Overall, I find some support for H1a. I also fail to find evidence of a significant interaction between *Evaluation Purpose* and *Rating Type* ($F_{2,138}$=0.518, p=0.597). Thus, I fail to find support for H2 with this alternative measure of *Narrative Evaluation Favorability*. Overall, across all 3 measure of *Narrative Evaluation Favorability* I find partial evidence of H1a and no evidence supporting H1b, H2a, nor H2b.

1.4.3.2 Process Measures

Although I generally did not find evidence supporting my hypotheses, it is possible that the independent variables in this study affected the supervisor goal/motivations predicted by

theory. I ask participants to indicate how motivated they were to be nice and honest in each piece of their performance evaluation on a scale from 0 (Not at all motivated) to 10 (Completely motivated). First, I ask them about their motivations for their narrative evaluation, then if applicable, I ask them about their motivations for their rating evaluation. These questions provide some insight about the use of each piece of the performance evaluation to satisfy honesty and social cost reduction goals. Since I expect the supervisor's motivation to satisfy these goals to differ by condition, I perform two separate untabulated ANOVA on the two motivation questions for the narrative evaluations with *Evaluation Purpose*, *Rating Type*, and the interaction between the two as factors. I fail to find any significant effects of the factors on either motivation to be honest or motivation to be nice in the narrative (all p's > 0.28). I also perform two untabulated ANOVA on the two motivation questions for the ratings with *Evaluation Purpose* as the factor. I fail to find evidence that the motivation to achieve either the goal to be nice or the goal to be honest in the ratings differ between conditions (all p's > 0.62). These results provide some evidence that the manipulations in the study failed to affect the motivation to accomplish the goals predicted in the study.

I also investigate whether *Narrative Evaluation Favorability* is related to the participants' reported motivations to be nice or honest in the narrative evaluation. Using untabulated simple linear regression, I find evidence that the motivations to be nice and honest in the narrative evaluation in the study are significantly related to *Narrative Evaluation Favorability* (using the *Narrative Evaluation Favorability* measure from Table 1). Specifically, motivation to be nice is positively related with *Narrative Evaluation Favorability* ($t_{df=142}$=3.731, two-tailed p<.01), and motivation to be honest is negatively related with *Narrative Evaluation Favorability* ($t_{df=142}$=-1.480, one-tailed p=.071). These findings are consistent with my expectations that supervisors make evaluations consistent with their goals even if those goals are not to be honest.

1.4.3.3 Understanding the Relationship Between Ratings and Narratives

For participants in conditions that require both a rating and a narrative, I ask multiple questions to capture the perspectives of participants on the relationship between ratings and narratives. They respond with their agreement on an 11-point scale (-5-Completely Disagree to 5-Completely Agree) to the following four questions: (1) "I felt the need to make my rating and written evaluation consistent with each other", (2) "My rating and written evaluation were different from each other", (3) "Providing a rating gave me the ability to be more honest in my written evaluation", and (4) "Providing a written evaluation gave me the ability to be more honest in my rating." On average, participants reported they felt the need to make the rating and narrative evaluation consistent with each other (mean=3.2, μ_0=0, $t_{df=69}$=11.451, two-tailed p<.01), their rating and written evaluation were not different from each other (mean=-2.143, μ_0=0, $t_{df=69}$=-6.405, two-tailed p<.01), they felt that providing the rating gave them the ability to be more honest in the narrative evaluation (mean=2.043, μ_0=0, $t_{df=69}$=7.713, two-tailed p<.01), and they felt that providing the written evaluation gave them the ability to be more honest in the rating (mean=2.229, μ_0=0, $t_{df=69}$=7.683, two-tailed p<.01). Overall, the responses to question (1) and (2) suggest that participants tend to match their ratings and narrative evaluations consistent with the theory presented for H2b. Although the responses to questions (3) and (4) seem consistent with some form of psychological licensing as suggested in H2a, the responses to question (1) and (2) make it less likely that participants are using narrative evaluations to achieve one goal and ratings to achieve another goal as predicted by strategic psychological licensing.

I continue by analyzing whether the responses to these four questions differ by experimental condition. In an untabulated regression, I find that responses to question (1) differ by *Evaluation Purpose* with participants in the coaching condition (relative to the bonus condition) reporting less agreement with the statement that they felt the need to make the rating and narrative consistent with each other (difference=-1.493, $t_{df=67}$=-2.253, two-tailed p=0.0275). Similarly, in another untabulated regression, I find that responses to question (2) differ by

Evaluation Purpose with participants in the coaching condition (relative to the bonus condition) reporting more agreement with the statement that their rating and narrative evaluation were different from each other (difference=2.113, $t_{df=67}$=2.677, two-tailed p=0.009). Responses to question (3) and (4) were not significantly different by condition (all p's > 0.33). The responses to question (1) and (2) provide some evidence that the perceived need to match ratings and narrative evaluations is stronger when compensation contracting (e.g., a bonus) is affected by the evaluation rather than when the evaluation is only used for coaching.

At the end of the experiment, I asked participants three more questions to understand the order they filled out the parts of the evaluation, their perceptions of the relative importance of the rating and narrative, and the relative thought and effort they put into the rating and narrative. Participants responded on 11-point scales with labels of 0-Written evaluation first (or written evaluation is clearly more important or 100% written evaluation effort) and 10-Numerical Rating first (or numerical rating is clearly more important or 100% numerical rating effort). On average, participants stated they filled out the rating before the narrative evaluation (mean=7.057, μ_0=5, $t_{df=69}$=4.447, two-tailed p<.01), believed the narrative was more important than the rating (mean=2.471, μ_0=5, $t_{df=69}$=-8.664, two-tailed p<.01), and put more thought and effort into the narrative evaluation compared to the rating (mean=2.157, μ_0=5, $t_{df=69}$=-12.177, two-tailed p<.01). This evidence is consistent with expectations that people believe the narrative evaluations are very important (Smither and Walker 2004). Since the design of this study places the numerical rating above the narrative evaluation on the page, I expected participants to fill out the rating first. This allowed me to focus my study on the impact of ratings on narratives rather than the other way around.

1.4.3.4 Analysis of Ratings

In the experiment, I also captured three additional conditions (72 additional participants from the same population) that were not discussed in the primary analysis. In these conditions, participants provide a rating but no narrative evaluation and are randomly assigned to be in one of

the three *Evaluation Purpose* conditions. Using these 3 additional conditions along with the three "with rating" conditions that also provide narrative evaluations, I run an ANOVA on the supervisors' ratings with *Evaluation Purpose*, whether a narrative evaluation is required or not, and the interaction between these two variables as factors. I find no significant effects (untabulated, all p's > 0.158). Although this study was not focused on establishing whether the presence of narrative evaluations affect numerical ratings, I note that I fail to find evidence that narrative evaluations influence ratings. However, these findings should be taken with caution as I expected the *Evaluation Purpose* variable to influence numerical ratings but failed to find evidence of an *Evaluation Purpose* effect. I expected to find ratings to be higher when numerical ratings are used in the control system (either for coaching or bonus purposes) compared to when the rating has no purpose and is kept private.

1.5 Conclusions

Understanding the effects of ratings on narrative evaluations is particularly important and timely given the push to remove subjective performance ratings in practice. Using an experiment, I investigate the effects of ratings and the purpose of evaluations on narrative evaluations using a laboratory experiment. I fail to find evidence for much of my theory. Although I do find some evidence that the presence of social costs from either an employee feedback or compensation contracting purpose do increase the favorability of narrative evaluations relative to a private evaluation (consistent with leniency effects due to social costs), I fail to find evidence that ratings affect narrative evaluations. Specifically, I fail to find evidence of the predicted interaction between ratings and the evaluation purpose on narrative evaluation favorability. Using theories on goal attainment in performance appraisal, psychological licensing, and process accountability, I predicted supervisors would attempt to satisfy goals of honesty and social cost reduction through the means available to them. I built theory suggesting that the way in which managers seek to satisfy those goals with narrative evaluations depends on whether they also give a rating and what the evaluation is going to be used for. When the evaluation is just for coaching, I expected to find

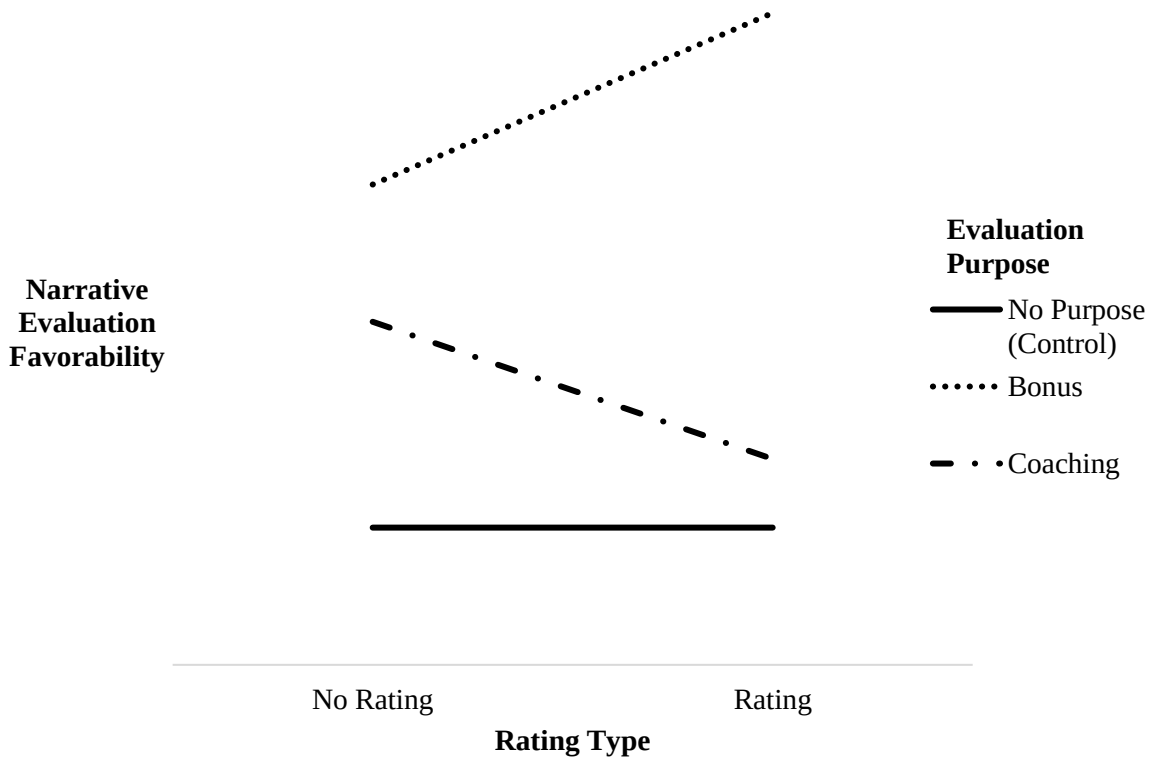
evidence of supervisors seeking a psychological license to free up their ability to be honest in narrative evaluations thereby satisfying both honesty and social cost reduction goals. Due to additional process accountability and the unambiguous nature of ratings, I predicted the presence of rating would increase leniency in narrative evaluations when the evaluation is used to administer a bonus. However, I failed to find evidence supporting these theories in this experiment.

This study identifies some challenges with investigating this research question that I attempt to overcome in Chapter 2 of this dissertation. The first challenge identified is cleanly manipulating the *Evaluation Purpose* while effectively helping participants understand the manipulation. The second challenge for this study is the noise inherent in a qualitative dependent variable like narrative evaluations. Upon visual inspection of the individual narrative evaluations, I note that the participants had vastly different methods of completing the narrative evaluations. This heterogeneity is understandable given the large spread of ages in my sample (min=21 years old, max=62 years old). Since narrative evaluations are inherently noisy, minimizing individual differences is critical to reduce the error variance sufficiently to detect differences between conditions. Noise in the dependent variable is likely to be one of the reasons for the failure to detect statistically significant results in this study.

In Chapter 2, I present a follow-up study that seeks to address these two major concerns. The goal of this follow-up study is still to understand the effects of ratings on narrative evaluations. To address the issues with *Evaluation Purpose*, I replace the *Evaluation Purpose* variable with a directional goal manipulation that is easier for participants to understand but still provides a goal for them to achieve which I expect to affect judgments. To address the issues with heterogeneity of participants causing insurmountable error variance, I use a more homogenous population of undergraduate business students as participants.

1.6 Figures and Tables for Chapter 1

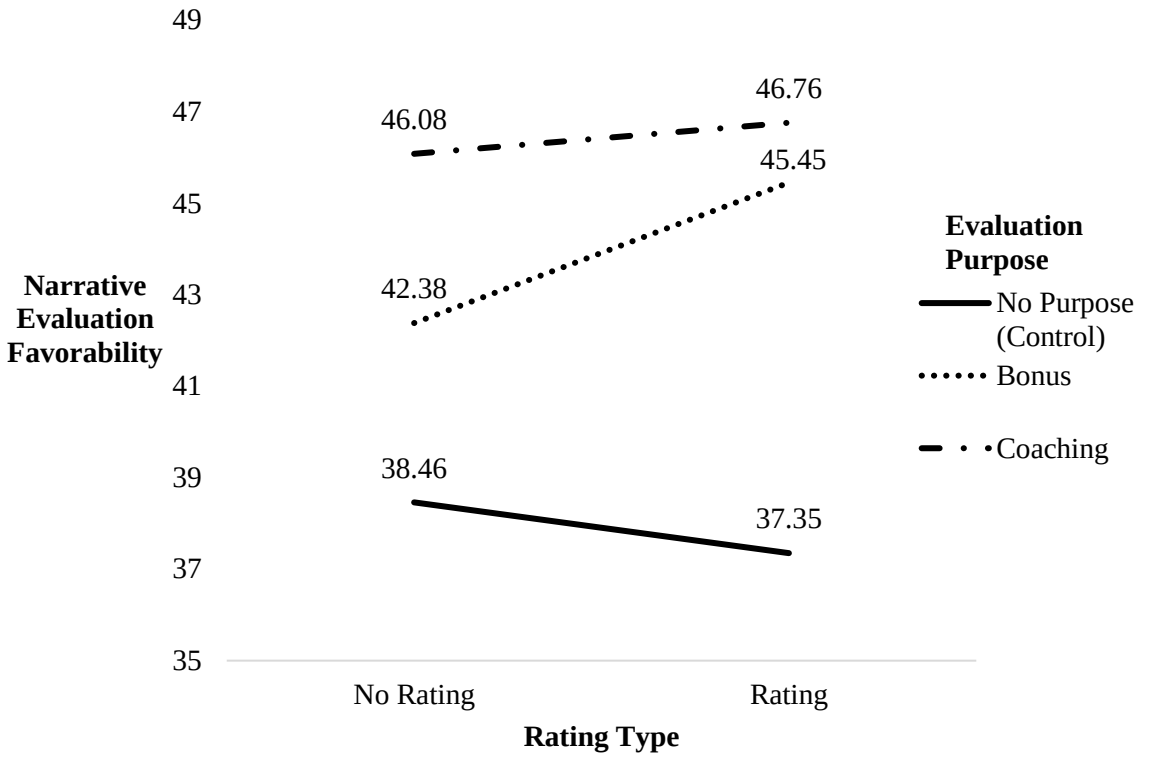
FIGURE 1 - Graphical Representation of Hypotheses 1 and 2



This graph displays the predicted pattern for *Narrative Evaluation Favorability* by experimental condition. All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Narrative Evaluation Favorability*, captures the overall favorability of the narrative evaluation provided by the supervisor. *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*Rating*) of employee performance while other did not provide a rating (*No Rating*). *Evaluation Purpose* was manipulated as *No Purpose, Bonus,* and *Coaching*. In the *Bonus* condition, the evaluation is provided by the supervisor was sent to a boss and used to decide whether the employee received a bonus. In the *Coaching* condition, the evaluation is only provided to the employee and is used just for coaching. In the *No Purpose* condition, the evaluation is kept private and not shown to the boss or the employee.

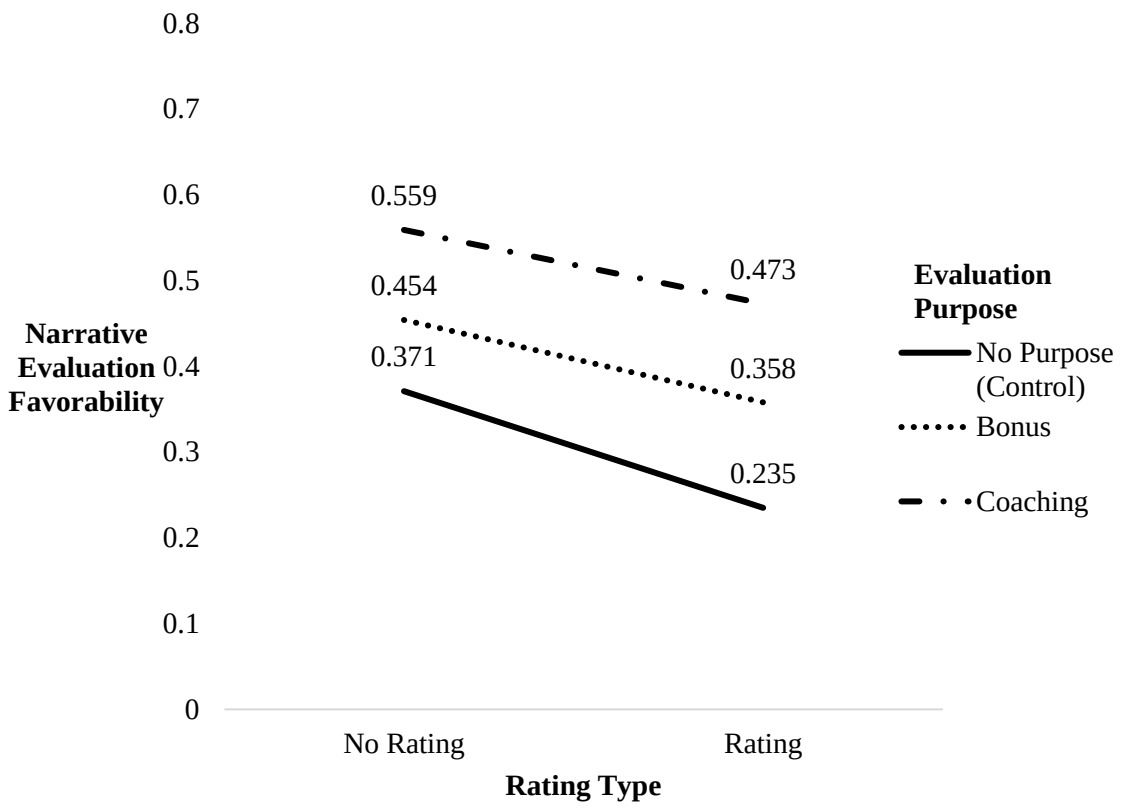
Specific Predictions
H1a predicts that the two *No Rating* conditions will be higher than the *No Purpose* (Control) condition.
H1b predicts that the *No Rating/Bonus* condition will be higher than the *No Rating/Coaching* condition.
H2a predicts that the *Rating/Coaching* condition will be lower than the *No Rating/Coaching* condition.
H2b predicts that the *Rating/Bonus* condition will be higher than the *No Rating/Bonus* condition.

FIGURE 2 - Graph of Cell Means for *Narrative Evaluation Favorability* as Manually Coded on 101-Point Scale



This graph displays the means for *Narrative Evaluation Favorability* by experimental condition (See Table 1). All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Narrative Evaluation Favorability*, is a measure of the favorability of each supervisor's narrative performance evaluation as coded by the author on a scale from 0 to 100 with 0 labeled as "Far Below Expectations" and 100 as "Far Above Expectations". *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*Rating*) of employee performance while other did not provide a rating (*No Rating*). *Evaluation Purpose* was manipulated as *No Purpose, Bonus,* and *Coaching*. In the *Bonus* condition, the evaluation is provided by the supervisor was sent to a boss and used to decide whether the employee received a bonus. In the *Coaching* condition, the evaluation is only provided to the employee and is used just for coaching. In the *No Purpose* condition, the evaluation is kept private and not shown to the boss or the employee.

FIGURE 3 - Graph of Cell Means for *Narrative Evaluation Favorability* Coded as the Sentiment Score Provided by IBM Watson Software



This graph displays the means for *Narrative Evaluation Favorability* by experimental condition (See Table 2). All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Narrative Evaluation Favorability*, is a measure of the favorability of each supervisor's narrative performance evaluation as coded using the Sentiment score provided by the Natural Language Understanding software from IBM Watson. *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*Rating*) of employee performance while other did not provide a rating (*No Rating*). *Evaluation Purpose* was manipulated as *No Purpose, Bonus,* and *Coaching*. In the *Bonus* condition, the evaluation is provided by the supervisor was sent to a boss and used to decide whether the employee received a bonus. In the *Coaching* condition, the evaluation is only provided to the employee and is used just for coaching. In the *No Purpose* condition, the evaluation is kept private and not shown to the boss or the employee.

TABLE 1 – Analysis of Narrative Evaluation Favorability as Manually Coded on 101-Point Scale

Panel A: *Narrative Evaluation Favorability* Mean (Standard Deviation) [Sample Size]

	Rating Type		
Evaluation Purpose	No Rating	Rating	Overall
No Purpose	38.46	37.35	37.91
	(11.10)	(9.43)	(10.22)
	[24]	[23]	[47]
Bonus	42.38	45.45	43.79
	(8.12)	(14.94)	(11.71)
	[26]	[22]	[48]
Coaching	46.08	46.76	46.43
	(16.20)	(13.61)	(14.78)
	[24]	[25]	[49]
Overall	42.31	43.26	42.77
	(12.39)	(13.35)	(12.83)
	[74]	[70]	[144]

Panel B: Two-way ANOVA Model of *Narrative Evaluation Favorability*

Source of Variation	SS	df	F	p-value
Intercept	262382	1	1675.928	<.01
Evaluation Purpose	1829	2	5.840	<.01
Rating Type	28	1	0.177	0.675
Evaluation Purpose X *Rating Type*	104	2	0.333	0.718
Error	21605	138		

Panel C: Follow-up simple effects tests to investigate H1a and H1b

Simple Effect	Estimate	$t_{df=138}$	p
Coaching vs. No Purpose (Without Rating)	7.625	2.111	0.037
Bonus vs. No Purpose (Without Rating)	3.926	1.109	0.270
Bonus vs. Coaching (Without Rating)	-3.699	-1.044	0.298

Table Notes - All p-values listed in the table are two-tailed. This table presents descriptive statistics, an ANOVA model, and simple effects for *Narrative Evaluation Favorability*. All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Narrative Evaluation Favorability*, is a measure of the favorability of each supervisor's narrative performance evaluation as coded by the author on a scale from 0 to 100 with 0 labeled as "Far Below Expectations" and 100 as "Far Above Expectations". *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*with rating*) of employee performance while other did not provide a rating (*no rating*). *Evaluation Purpose* was manipulated as *No Purpose*, *Bonus*, and *Coaching*. In the *Bonus* condition, the evaluation is provided by the supervisor was sent to a boss and used to decide whether the employee received a bonus. In the *Coaching* condition, the evaluation is only provided to the employee and is used just for coaching. In the *No Purpose* condition, the evaluation is kept private and not shown to the boss or the employee.

TABLE 2 – Analysis of Narrative Evaluation Favorability Coded as the Sentiment Score Provided by IBM Watson Software

Panel A: *Narrative Evaluation Favorability* Mean (Standard Deviation) [Sample Size]

	Rating Type		
Evaluation Purpose	No Rating	Rating	Overall
No Purpose	0.371	0.235	0.304
	(0.457)	(0.553)	(0.505)
	[24]	[23]	[47]
Bonus	0.454	0.358	0.410
	(0.423)	(0.507)	(0.461)
	[26]	[22]	[48]
Coaching	0.559	0.473	0.515
	(0.385)	(0.440)	(0.412)
	[24]	[25]	[49]
Overall	0.461	0.359	0.411
	(0.424)	(0.503)	(0.465)
	[74]	[70]	[144]

Panel B: Two-way ANOVA Model of *Narrative Evaluation Favorability*

Source of Variation	SS	df	F	p-value
Intercept	23.945	1	112.145	<.01
Evaluation Purpose	1.091	2	2.555	0.081
Rating Type	0.401	1	1.876	0.173
Evaluation Purpose X *Rating Type*	0.016	2	0.038	0.963
Error	29.465	138		

Panel C: Follow-up simple effects tests to investigate H1a and H1b

Simple Effect	Estimate	$t_{df=138}$	p
Coaching vs. No Purpose (Without Rating)	0.189	1.413	0.160
Bonus vs. No Purpose (Without Rating)	0.084	0.639	0.524
Bonus vs. Coaching (Without Rating)	-0.105	-0.802	0.424

Table Notes - All p-values listed in the table are two-tailed. This table presents descriptive statistics, an ANOVA model, and simple effects for *Narrative Evaluation Favorability*. All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Narrative Evaluation Favorability*, is a measure of the favorability of each supervisor's narrative performance evaluation as coded using the Sentiment score provided by the Natural Language Understanding software from IBM Watson. Each observation receives a score from -1 to 1 with higher numbers representing more positive word content. *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*with rating*) of employee performance while other did not provide a rating (*no rating*). *Evaluation Purpose* was manipulated as *No Purpose, Bonus,* and *Coaching*. In the *Bonus* condition, the evaluation is provided by the supervisor was sent to a boss and used to decide whether the employee received a bonus. In the *Coaching* condition, the evaluation is only provided to the employee and is used just for coaching. In the *No Purpose* condition, the evaluation is kept private and not shown to the boss or the employee.

CHAPTER 2

THE EFFECTS OF RATINGS AND DIRECTIONAL GOALS ON NARRATIVE EVALUATIONS

2.1 Introduction

In the recent upheaval of traditional performance appraisal systems, some high-profile companies have removed numerical performance ratings from their appraisal process (Rock and Jones 2015; Silverman 2016). While the benefits of this removal are questioned by some proponents of ratings because they believe ratings are useful despite having faults (Adler et al. 2016), we don't know how ratings (or the lack of ratings) impact the narrative evaluations commonly provided to employees as part of the management control system. This practical dilemma facing companies provokes an interesting question. How does quantifying a subjective performance evaluation with a rating affect detailed subjective narrative evaluations that traditionally accompany the ratings? Previous research on performance evaluation focuses almost exclusively on the performance rating as the outcome of interest, leaving the narrative evaluation under-researched (Wilson 2010; Speer 2018). I perform a controlled experiment to fill this gap in practical and theoretical knowledge regarding the effects of quantifying a subjective performance evaluation on narrative performance evaluations.

Subjective performance evaluations provide a way for firms to measure, store, and use information about employee performance that is difficult to objectively measure. Subjectively interpreting raw performance information into summarized evaluations provides information for use in the firms' management control systems. Although numerical subjective performance measures have historically played a significant role in determining important outcomes like the administration of compensation contracts (Bol 2008), these ratings are not without issues. At times, these ratings have been shown to be influenced by contextual features common in business environments (e.g., knowledge of employee performance on different tasks, see Bol and Smith 2011). The effects of these contextual features (often, resulting in biased evaluations) and

concerns about the general usefulness of performance ratings are at the center of the debate over getting rid of performance ratings (Adler et al. 2016). The collective knowledge of the content of narrative evaluations and potential bias in these narratives is limited (Wilson 2010; Speer 2018). As formal evaluations affect compensation and career outcomes for employees and provide feedback to employees on their performance, bias in performance evaluations can impact a firm's ability to maximize employee motivation, satisfaction, and performance (Ittner, Larcker, and Meyer 2003).

Previous research on performance ratings asserts that supervisors often have implicit or explicit goals that lead them away from accurate evaluations (Murphy and Cleveland 1995). When supervisors have directional goals to either give a higher or lower evaluation to an employee, motivated reasoning theory predicts they will attempt to rationalize biased behavior while maintaining a belief they are being objective (Kunda 1990). I expect this motivated reasoning process to affect the overall favorability of narrative evaluations with positive (negative) directional goals leading to more (less) favorable narrative evaluations than is warranted by actual employee performance. In other words, I expect narrative evaluations to be impacted by the directional goals.

In isolation (e.g., in evaluations without numerical ratings), motivated reasoning theory predicts narrative evaluations are biased by directional goals. However, research comparing numbers (with context-specific labels to provide meaning) and words suggests that use of number and words to communicate ideas may be affected differently by directional goals (Daft and Wiginton 1979; Piercey 2009). As a lower variety language, numerical rating scales are highly context-specific and have a limited number of interpretations. As a higher variety language, general verbal expression can be used for a variety of situations which makes it more ambiguous in its meaning (i.e., has a larger number of possible interpretations) (Daft and Wiginton 1979). This reduced ambiguity and limited freedom of interpretation for numerical rating scales reduces its susceptibility to motivated reasoning in the presence of directional goals (Piercey 2009). In

other words, the lack of "wiggle room" when interpreting the meaning of numerical ratings makes it relatively difficult to give a preference-consistent rating and still maintain the necessary "illusion of objectivity" (Kunda 1990). Since I expect the ratings supervisors provide to be more resistant to the effects of directional goals, supervisors' narrative evaluations will be constrained by these relatively objective ratings due to their perceived need to match the message of the rating with the narrative evaluation (David 2013; Erdogan 2002; Kunda 1990). The end result is a less preference-consistent narrative evaluation when the supervisor also has to provide a numerical rating.

To address my research question and test these expectations, I perform a 2x2 between-subjects experiment manipulating rating type and directional goal where undergraduate business student participants take the role of a supervisor at a hypothetical company and evaluate an employee's performance. After each participant views a video of the employee's performance, they must fill out a performance evaluation for the employee. Since I am focused on understanding how narrative evaluations change under different conditions, all participants provide a narrative evaluation. The rating type is manipulated such that the evaluation either requires the participant to assign a numerical rating to the employee's performance or does not require a rating. I manipulate directional goal as either a positive or negative evaluation goal by providing advice from the participant's hypothetical boss. In the positive (negative) goal, the boss describes historical evaluation behavior in the company as being too harsh (nice) and requests that the participant be honest but realize the need to be more nice (harsh). The dependent variable of interest is the *Narrative Evaluation Favorability* measured using the average of two independent coders' judgments of the favorability of each narrative evaluation.

In my experiment, I find evidence that the favorability of narrative evaluations is affected by directional goals. Also, the effect of directional goals on narrative evaluation favorability is reduced when supervisors must also provide a numerical rating. This evidence is consistent with theory that supervisors write narrative performance evaluations that attempt to achieve their

directional goal due to motivated reasoning, but the act of rating an employee on a scale constrains that motivated reasoning in the narrative evaluation. These findings suggest that firms using ratingless evaluations may have narrative evaluations more affected by supervisors' directional goals than firms with narrative evaluations that also require numerical ratings.

This study contributes to the heated debate on the removal of performance ratings (Adler et al. 2016; CEB 2016; Cappelli and Tavis 2016) by providing nuanced insight into one potential impact of this trend. First, I show that narrative evaluations are affected by directional goals. So, justifying removal of ratings due to inaccuracy may be unjustified as the narrative evaluations are also affected by evaluator preferences. In addition, I show the traditional method of providing a rating along with the narrative evaluation reduces the effects of directional goals on narrative evaluations. This insight should be useful to practitioners weighing options for changing their performance evaluation process.

I also contribute to the research stream investigating subjective performance evaluation by deepening the collective understanding of factors that affect narrative evaluations. Subjective narrative evaluations have received very little attention in the performance evaluation literature (Wilson 2010; Speer 2018). This study provides evidence that narrative evaluations are not simply qualitative representations of their quantitative counterparts. Instead, they are affected by the process of giving ratings. This study also contributes specifically to a management accounting audience because it investigates how subjective measurement by supervisors impacts their qualitative evaluation of the employee. While a recent study in management accounting investigates how qualitative feedback information (e.g., causal words) impacts the recipient (Loftus and Tanlu 2018), my study is the first study in accounting to my knowledge that investigates the determinants of subjective narrative information in the control system.

I also extend the management accounting literature on performance feedback more generally by looking into determinants of performance feedback rather than outcomes of feedback. While substantial literature has focused on the effects of objective control system

feedback on employee effort (e.g., Hannan, Krishnan, and Newman 2008; Hannan, McPhee, Newman, and Tafkov 2013; Casas-Arce, Lourenço, and Asís Martínez-Jerez 2017), very little has investigated how managers determine what performance feedback to give.

Last, I contribute to the broader accounting literature on narrative communication (e.g., Li 2008; Asay, Libby, and Rennekamp 2018; Bentley 2019). These studies investigate how agents communicate with principals using narrative reports and how it affects the agents and principals. I extend the literature by investigating how principals communicate using narrative reports to agents. Specifically, I investigate how principals' narrative evaluations of agents are affected by a need to assign quantitative ratings.

This paper continues with Section II establishing the background and developing the hypotheses. Section III and Section IV describe the methodology used to investigate the research question and the results of the study, respectively. Section V concludes the paper.

2.2 Background and Hypotheses

2.2.1 Background on Subjective Performance Evaluation

"Management accounting systems facilitate decision making by collecting and reporting relevant information for alternatives being considered" (Swieringa and Weick 1981). Guided by the value-added principle, firms measure and use information that helps them run a successful business (Edmonds, Edmonds, Edmonds, Edmonds, and Olds 2020). Measuring and evaluating the performance of the firm, business units, projects, and employees are a critical purpose of management accounting systems. Managing employee performance is on the minds of leaders in prominent accounting firms like Deloitte (Buckingham and Goodall 2015). This study focuses specifically on evaluating employee performance. To manage employee performance, firms need to measure performance and compare it to expectations. When it is difficult to objectively measure performance, subjective evaluations are often used to provide essential information to the management control system. This subjective performance evaluation fills in gaps left by objective measures to achieve the multiple roles of control systems (Demski and Feltham 1976).

Subjective performance evaluation plays a role in compensation contracting through its use in performance-related decisions like promotion, bonus allocation, and raise determination (Bol 2008). Research in management accounting has provided important insights into the use of subjectivity in these types of compensation contracting decisions (e.g., Bol 2011; Bol and Smith 2011; Chen, Jermias, and Panggabean 2016).

Subjective performance evaluation also fills a feedback role in control systems by providing crucial information on employee performance to the employee for development and to better understand how their performance compares to firm expectations (London 2003). The feedback literature in management accounting largely focuses on how different characteristics of objective feedback (e.g., feedback not influenced by supervisor subjectivity) impact employee performance. Recent studies investigating this objective feedback include those on relative performance information (Hannan et al. 2008; Hannan et al. 2013) and absolute performance information (Casas-Arce et al. 2017). While these studies emphasize the importance of feedback in management accounting, they focus on information that is not subjectively generated by supervisors. However, performance that is difficult to measure objectively will still need to be subjectively evaluated by supervisors to give employees the feedback they need and serve as the basis of compensation contracting decisions. Thus, understanding determinants of subjective feedback is important because of its effects on firm decision-making (e.g., compensation contracting) and employee development.

While the past literature on subjective performance evaluation has focused almost exclusively on numerical performance information, narrative or verbal evaluations play an important role in control systems. In traditional performance appraisal used for compensation contracting, supervisors commonly provide both performance ratings and narrative evaluations of employee performance (Gorman, Meriac, Roch, Ray, and Gamble 2017). The content of these narrative summaries remains a new and under-researched area of performance evaluation (Speer 2018; Wilson 2010), even though employees report paying more attention to the narrative

evaluation than the numerical rating (Smither and Walker 2004). Narrative performance evaluations allow for richness and nuance unavailable using performance ratings that can be useful to employees and decisionmakers. Given the importance of narrative evaluations in practice and the recent removals of performance ratings mentioned in the introduction of this paper, understanding how narrative evaluations change with and without performance ratings is a timely and important question.

2.2.2 Issues with Subjective Performance Evaluation

To understand how performance ratings might impact narrative performance evaluations, I rely on the goal attainment framework of subjective performance evaluation presented by Murphy and Cleveland (1995). One of the important tenets of this framework is that supervisors can have goals besides being perfectly accurate in their evaluations. These goals can result in biases. In the accounting literature, a variety of biases for numerical performance ratings have been identified including: leniency and compression (Moers 2005), previous knowledge of employee performance (Bol and Smith 2011; Tan and Jamal 2001), part-time status (White 2019), employee's previous evaluation (Woods 2012), inconsistent use of subjectivity in balanced scorecard evaluations (Ittner et al. 2003), and favoritism (Du, Tang, and Young 2012).

Biased performance information is a concern for both compensation contracting and employee feedback in the firm. For compensation contracting, biased evaluations make it more difficult to fairly distribute contracted rewards and make other personnel decisions like promotion, which can impact employee motivation negatively (Bol, Kramer, and Maas 2016). For feedback, biased evaluations compromise opportunities for employees to develop and learn. Since performance feedback information has long since been considered useful for performance improvement (Kluger and DiNisi 1996), poor quality performance information has the potential to restrict employee development and success.

The investigation of performance evaluation biases in accounting has focused exclusively on numerical performance ratings. These studies focus on the construct of rating favorability as

measured by how high or how low the evaluation is on a scale for a given employee. Alternative constructs of importance in the performance evaluation context include internal consistency and interrater agreement, among others (Murphy and Cleveland 1995). To maintain comparability to ratings and the previous literature, my construct of interest in this study is the overall favorability of the narrative content[10]. Narrative evaluation favorability is the holistic picture (based on the narrative) of how well the individual performed, with more favorable evaluations indicating better performance and less favorable evaluations indicating worse performance. In this study, I focus on building theory for factors that affect narrative evaluation favorability.

2.2.3 Motivated Reasoning

Murphy and Cleveland (1995) suggest that inaccurate ratings are not necessarily evidence of an inability to accurately judge performance due to limited understanding of the performance context or lack of ability to encode and understand the information. Instead, inaccurate ratings are evidence that supervisors have goals other than being perfectly accurate in evaluations. If someone's goals push them away from accuracy or neutrality, the goals can be considered directional on either side of accurate. Directional goals are goals either explicitly or implicitly held by a person that motivate a person "to arrive at a particular, directional conclusion" (Kunda 1990). In accounting settings, a variety of directional goals exist and have been shown to affect judgments. Examples include directional goals to make audit judgments consistent with a partner's preference (Wilks 2002; Piercey 2009; Peecher, Piercey, Rich, and Tubbs 2010), to make money due to increasing or decreasing investment value based on investment position (Hales 2007), to make forecasts consistent with incentives (Bradshaw, Lee, and Peterson 2016), and to be viewed as competent by themselves and peers (Tayler 2010). In a performance evaluation setting, directional goals can take the form of a directional preference to provide an

[10] Other potentially interesting constructs for detecting variation in narrative evaluations include accuracy of events recalled and noted, completeness of narrative, interrater reliability, test-retest reliability, among others.

evaluation describing a specific level of employee performance. For example, a supervisor might have a vested interest in the success of an employee that the supervisor personally hired. When the time comes to evaluate the employee's performance, the supervisor may have an implicit or explicit goal to evaluate this employee as having high performance to manage perceptions of the supervisor's ability to hire and mentor employees. Similarly, supervisors can have negative evaluation goals. For example, a supervisor may have an implicit or explicit goal to evaluate an employee negatively to confirm a previously held belief about the employee stemming from a stereotype (e.g., race, gender).

To understand the potential impact of directional goals on narrative evaluation favorability, I appeal to the literature on motivated reasoning. Kunda (1990) suggests that individuals with directional goals will attempt to reason their way to judgments that are consistent with their desired conclusion. Thus, in the presence of relatively complex and ambiguous information like in a subjective performance evaluation scenario, individuals will have enough leeway to reach their desired conclusion through rational justification. This allows them to "maintain an illusion of objectivity" (Kunda 1990) while still achieving their directional goal. In other words, they believe they are being accurate and unbiased even when they are not. Research finds this rationalization process is largely unconscious making it particularly worrisome for decision quality (Wilks 2002; Piercey 2009). Given a directional goal to evaluate an employee positively (negatively), I expect the favorability of the supervisor's narrative evaluation to be more favorable (less favorable), thereby attempting to satisfy the supervisor's directional goal.

H1: The favorability of supervisors' narrative evaluations is higher when they have a positive directional goal relative to a negative directional goal.

While H1 considers whether the narrative evaluations themselves are affected by directional goals, it does not consider the impact of ratings on the favorability of narrative evaluations. The following section builds theory on how the act of rating an employee's

performance on a numerical scale may impact the directional preference effects present in narrative evaluations.

2.2.4 Reasonableness Constraints – Numerical Rating

To this point, I have focused on the presence of motivated reasoning in narrative evaluations. However, traditional performance evaluations include both a numerical rating and a narrative evaluation (Gorman et al. 2017). Thus, it is important to consider how motivated reasoning might affect the favorability of ratings and narrative evaluations differently. Daft and Wiginton (1979) suggest that communication methods in organizations differ along the dimension of language variety with the spectrum of language variety ranging from high to low. High variety language (e.g., art and body language) is characterized by its ambiguity which allows it to express a great variety of complex concepts, but at the same time allows it to be interpreted in many ways. Low variety language (e.g., analytical mathematics) is more precise which makes the specific language useful in only a narrow set of circumstances. This language specificity also reduces the number of possible interpretations. Along Daft and Wiginton's (1979) proposed spectrum, "general verbal expression" is considered a higher variety language than "linguistic variable (semantic differential, Likert Scale)." Being a high variety language, general verbal expression has significant variety to communicate many things in many ways. This freedom of interpretation by both author and reader introduces ambiguity about what is meant by a specific verbal expression. Narrative evaluations use general verbal expression. A numerical rating which usually takes the form of a scale with meaning derived from scale labels is a relatively low variety language. Since rating scales are a lower variety language, there is less room for either the rater or the ratee to misinterpret the meaning of a given numerical rating. This concept has been specifically applied to performance evaluation ratings and narratives in a thought piece by Brutus (2010). Brutus provided some initial theoretical footing for why we might expect ratees and raters to be affected by the format of the evaluations (rating portion

versus narrative portion). However, Brutus (2010) only considers the context of simultaneously presented ratings and narratives. So, Brutus does not consider how narratives might differ with and without ratings. In my study, I compare a narrative without a rating to a narrative with a rating.

The effects of ambiguity on decision making has been investigated in multiple accounting contexts with the general finding being that increased ambiguity in decision-relevant information allows greater flexibility to reach desired conclusions. Studies find that vagueness in financial accounting standards and tax law can lead auditors and tax professionals to take more aggressive, client-supporting positions (Hackenbrack and Nelson 1996; Spilker, Worsham, and Prawitt 1999). Similar findings in investor decision-making show that investors make more preference-consistent judgments when they receive management forecast guidance as a range rather than a point estimate. These studies describe how ambiguity in the information received by the decision maker affects decision-making. My study focuses instead on how the ambiguity of the information communication method (narrative versus rating) used by the decision-maker affects the decision makers own judgments. The accounting studies most similar to my study focus on the differences between verbal probability phrases (e.g., "probable") and numerical probabilities in accounting settings. These studies find that verbal probability statements commonly used in accounting standards (e.g., "remote") are vague and allow a variety of interpretations (Amer, Hackenbrack, and Nelson 1994, 1995; Cuccia, Hackenbrack, and Nelson 1995; Piercey 2009). While no differences in tax accountants' decisions were detected between a numerical and verbal probability standards in Cuccia, Hackenbrack, and Nelson (1995), they find evidence of accountants primarily using the flexibility in interpreting the meaning of the standard to achieve their goals when the standard used a verbal probability. One major difference between these previously discussed studies and my study is that these vague verbal judgments are being provided to participants to use to make judgments rather than the participants providing these verbal judgments themselves. One study, Piercey (2009), has participants provide verbal and

numerical probability judgments and assesses the tendency for these judgments to be impacted by directional goals. I describe the findings for this study and its relevant theory below.

In a performance evaluation setting, the reduced ambiguity of interpretation for ratings (relative to narratives) by both evaluator and evaluatee may impact the judgment's sensitivity to motivated reasoning. In a study comparing numerical and verbal probability judgments in an audit setting, Piercey (2009) presents experimental evidence that numerical probability judgments are less biased by directional goals than verbal probability judgments. He argues and provides evidence that the additional motivated reasoning for verbal probability judgments is driven by the ability for the decision maker to "re-define" the meaning of their judgment. This "re-definition" allowed the decision makers much more flexibility to make a biased judgment while maintaining an "illusion of objectivity" (Kunda 1990). This fits into the conceptual framework provided by Daft and Wiginton (1979) that number scales are lower variety communication methods than general verbal expression. In a performance evaluation setting, the meaning of a numerical rating is precise and permanent which reduces ambiguity of meaning. With limited ambiguity of meaning, these ratings are inelastic and difficult to use in the process of making a rational judgment consistent with a directional goal. Thus, motivated reasoning will be constrained for numerical ratings relative to narrative evaluations. Formally stated, the low variety (inelastic) nature of numerical ratings introduces a reasonableness constraint to the motivated reasoning process (Kunda 1990). In this setting, I expect the favorability of narrative evaluations (without a rating) to be more impacted by directional goals than will numerical ratings, resulting in less preference-consistent numerical ratings.

The title of this study poses the question, "Do ratings change the narrative?" I now build theory about how providing a less preference-consistent numerical rating (due to the lower language variety) will affect the susceptibility of narrative evaluations that accompany the ratings to directional goal effects. Evidence from performance evaluation research suggests that narrative evaluations can play the role of satisfying process accountability by justifying or explaining the

rating (David 2013; Erdogan 2002). In the presence of process accountability either to others or themselves (to maintain an illusion of objectivity), supervisors may feel pressure to match the favorability of the narrative evaluation with the ratings. An inconsistent message between the rating and the narrative may present salient evidence to the evaluator that they are being biased in one of the forms of evaluation which compromises the illusion of objectivity (Kunda 1990). Once the supervisor has determined to give a rating which will be relatively uninfluenced by directional goals, the rating plays the role of a reasonableness constraint on the narrative evaluation reducing the effect of directional goals on narrative evaluation favorability. Thus, I expect the favorability of the narrative evaluation that accompanies the rating to match the rating's favorability. This matching results in a narrative evaluation that is less affected by directional goals compared to a narrative evaluation without a rating. I predict the effect of directional goals on narrative evaluation favorability will be smaller when the supervisor must provide a numerical rating along with the narrative evaluation.

> *H2: The difference in narrative evaluation favorability between supervisors with a positive and a negative directional goal will be smaller when the supervisor is also required to give a numerical rating.*

2.3 Methodology

To test my hypothesis, I administer a 2x2 between-subjects experiment where participants act as supervisors in a hypothetical company and evaluate the performance of an employee[11]. I manipulate the participant's directional goal and whether the participant is required to rate the performance of the employee on a numerical scale. All participants provide narrative evaluations of the employee's performance.

[11] This study was approved by the Institutional Review Board at the author's institution.

2.3.1 Participants

Participants are 189 undergraduate business students at a large public university in the Northeastern United States[12]. 60.9% (38.6%) of participants state they are male (female)[13]. The median age of participants is 19 years old. I chose a participant pool that is appropriate for the task (Libby, Bloomfield, and Nelson 2002). In the task, they are required to evaluate the performance of a musician. All the participants report listening to music at least a couple of times a week with 75.6% of participants reporting they listen to music multiple times a day. Thus, these participants are well acquainted with music and can be expected to make reasonable judgments as to the quality of a musician's performance.

2.3.2 Procedures

Participants take the role of a supervisor (titled Cruise Director) in a hypothetical cruise ship vacation company. After being filled in on basic information about the company, the participant views images of a buffet line and gives narrative feedback on the quality of the work done by the dining staff[14]. Then, the participant is told they will evaluate an employee's (Tom) musical performance from a recent cruise. The participant then reads advice from their boss containing a directional goal for the performance evaluation, either positive or negative depending on the condition. After reading the form that the evaluation will take, the participant views a 3.5 minute video of Tom's musical performance. Then, the participant evaluates Tom using the performance evaluation form based on their condition, narrative evaluation with or

[12] Participants received course credit for participating in the research study from their instructor which was not the author.

[13] One participant preferred not to provide a response to the gender question.

[14] I included this narrative requirement before the primary task in all conditions to capture individual differences in narrative evaluation favorability to use as a covariate in the statistical models. Narrative favorability for the feedback to the dining staff was measured using the average judgment of two independent coders as described in the "Dependent Variable" section of this paper. As expected, the narrative favorability of the feedback given to the dining staff was positively related to the narrative evaluation favorability of the primary performance evaluation suggesting that individuals have tendencies to write more positively or negatively (μ=0.199, t_{184}=3.055, two-tailed $p<0.01$). When the narrative favorability for the feedback to the dining staff is included in the primary model (ANCOVA and simple effects), the results are qualitative similar and inferences remain the same. For ease of presentation and interpretation, I report the results in tables and text without controlling for these individual differences.

without a rating. Finally, participants answer post-experimental questions to help me better understand why they behaved the way they did and to collect demographic information.[15]

2.3.3 Task Performance Video

I used a video of a musical performance for specific experimental design purposes. First, a musical performance is considered an inherently subjective task. Since subjectivity is at the center of subjective performance evaluation and is important in the theory presented in this study, a highly subjective performance context like a musical performance is warranted. Second, real world performance evaluations are based on a rich set of information to draw on which allows for significant variation in narrative evaluations. Traditionally used experimental instruments for testing performance ratings are not sufficiently rich to allow for adequate variation in narrative evaluations. To balance the time constraints of participants and the richness of data to evaluate, I chose a task that is both brief and rich with data to evaluate. By viewing the performance of a single song, evaluators can focus on a variety of aspects: audience interaction, vocal quality, instrumentation quality, mistakes, stage presence, etc. This task provides an effective and efficient means of testing the hypotheses. Last, music is a task that many people are familiar with. It can be difficult to find a sufficient pool of participants that are able to write intelligently about and evaluate performance on a specific task. Given the ubiquity of music among university students, I was able to efficiently access participants that were reasonably equipped to write about an employee's musical performance.

2.3.4 Manipulations

2.3.4.1 Directional Goal

Directional goal is manipulated using advice from the participants' hypothetical boss. The theoretical construct of a directional goal is a supervisor either having a goal to be more positive or more negative in an evaluation. The advice from the boss exogenously provides

[15] Participants spent a median of about 26 minutes to complete the task.

supervisors with a directional goal in their evaluation. I follow prior directional goal research in auditing where auditors were instructed by their partner to be pessimistic (conservative) or optimistic (aggressive) due to client's preferences for favorable reporting (Piercey 2009; Peecher et al. 2010). To maintain believability in these audit studies, the audit partner provides contextually reasonable explanations for these preferences for pessimistic or optimistic judgments. Following this pattern, I manipulated directional goals in the following manner. The advice in all conditions begins by requesting that the performance evaluations be honest. Then, in the positive directional goal condition, the boss describes past experiences where the tendency of cruise directors has been to be too harsh in evaluations. The boss then recommends that the participant not be too harsh and to pay attention to things that go well in the performance. In the negative directional goal, the boss describes past experiences where the tendency of cruise directors has been to be too nice in evaluations. The boss then recommends that the participant not be too nice and to pay attention to things that do not go well in the performance. This directional goal manipulation gives participants a directional goal from their boss to be more positive or more negative in their evaluations of the employee's performance.

2.3.4.2 Rating Type

Rating type is manipulated at two levels. Participants in the no rating condition, *No Rating*, are not asked to provide a numerical rating in their performance evaluation. Participants in the rating present condition, *Rating*, are asked to rate the performance of the employee, Tom, on a scale based on their expectations for a professional performer at Riverboat Cruises Inc.[16] They respond on a scale with a number from 1 to 5 with labels of Unacceptable, Below Average, Average, Above Average, and Outstanding. In the *Rating* conditions, the performance rating scale

[16] While performance ratings can take the form of absolute or relative formats (like rankings), absolute ratings appear to be more commonly used in practice (Gorman et al. 2017). I use an absolute rating scale in this study for external validity and simplicity with supervisors only needing to evaluate a single employee.

is presented simultaneously with the narrative evaluation. The rating is at the top of the page and the narrative evaluation follows.

2.3.5 Dependent Variable

For narrative evaluations, participants responded to the following prompt, "Please provide a written evaluation of your employee, Tom, based on your expectations for a professional performer at Riverboat Cruises Inc [Please write a minimum of 250 characters]." Participants were forced to write at least 250 characters to ensure that the text responses would be long enough to allow for measurable variation in the construct of interest. I use the narrative evaluations as the raw material to create the dependent variable. My construct of interest is *Narrative Evaluation Favorability*. I measure this using the judgments of two independent coders.[17] Coders were given the following instructions, "Please rate each row separately as an independent evaluation. How favorable is the evaluation of Tom's music performance?" They rated the favorability of each narrative evaluation on a 1 to 7 Likert scale with 1 being Completely Unfavorable and 7 being Completely Favorable. Each coder viewed the narrative evaluations and provided a single judgment for each narrative evaluation. The two coders completed the assessments independently. Coders did not have access to the performance ratings provided by participants in the rating condition.[18] The coders showed high agreement. Specifically, the Pearson correlation coefficient is 0.909 (t_{187}=29.947, two-tailed p<.01). I also test the Intraclass Correlation of the two coders responses. I use the Intraclass Correlation model with the following assumptions: twoway random effects model, test for consistency (rather than

[17] The two coders were non-author doctoral students in accounting. The coders were blind to the experimental condition and made coding judgments of each participant's response in different random orders.

[18] It is possible that participants described a rating in the narrative evaluation. While I chose not to restrict the supervisors' ability to use any means they feel necessary in their narrative evaluations (including talking about ratings), it is possible that the mentioning of a rating in the narrative evaluation might unduly influence the coders evaluation of narrative evaluation favorability. Eight participants mentioned a variation of the word "rating" (e.g., rate, rating) or a number that could be conceived as a rating. I rerun the analyses in Table 1, Table 2, and Table 3 dropping these eight participants (untabulated). All results are qualitatively similar and inferences are unchanged.

absolute agreement), and the coding will be used for average measures. [19] The Intraclass Correlation Coefficient (ICC) is 0.95 with a 95% Confidence Interval for ICC of [0.934, 0.963]. This ICC is indicative of excellent reliability of the coders (Cicchetti 1994; Hallgren 2012). I average the responses by each of the coders to get the primary dependent variable for this study, *Narrative Evaluation Favorability*. I interpret higher values of *Narrative Evaluation Favorability* as more favorable. Please see the Appendix for two examples narrative evaluations provided by participants.

2.4 Results

2.4.1 Manipulation Check

Before analyzing the primary dependent variable, I assess the effectiveness of the *Directional Goal* manipulation by analyzing responses to the following two questions: (1) "How much pressure did you feel to be nice in your evaluation of Tom?" and (2) "How much pressure did you feel to be harsh in your evaluation of Tom?". Participants responded to each of these questions on 11-point scales with labels of "No Pressure At All" (0) to "Significant Pressure" (10). Participants in the positive directional goal conditions reported higher pressure to be nice (mean = 4.989) than participants in the negative directional goal conditions (mean = 4.118, difference=0.871, t_{187}=2.147, two-tailed p=0.033). Similarly, participants in the positive directional goal conditions reported less pressure to be harsh (mean = 2.958) than participants in the negative directional goal conditions (mean = 4.312, difference= -1.353, t_{187}= -3.598, two-tailed p<.01). These results suggest that the *Directional Goal* manipulation effectively manipulated positive and negative goals between participants.[20] While this data suggests the *Directional Goal* manipulation was successful on average, I also check whether each participant

[19] I use the "irr" package in R to calculate and statistically test the Intraclass Coefficient.
[20] I did not capture a manipulation check for the *Rating* manipulation. Unlike manipulations of theoretical constructs that represent psychological constructs that may need to be measured to ensure proper manipulation of subtle constructs, *Rating* is an important context-rich task that participants either complete or do not complete depending on their experimental condition. I manipulated ratings using a basic representation of what a rating would entail in practice.

attended to the directional goal manipulation. At the end of the experiment, I ask participants, "Which of the following quotes from your boss did you read?" They see the paragraph of advice from the boss in each condition and select which of the two they read. Only five participants answered this question regarding *Directional Goal* incorrectly suggesting most participants attended to the manipulation.[21]

2.4.2 Descriptive Statistics

The narrative evaluations produced by participants had a mean (median) number of words of 99.624 (83) consisting of a mean (median) of 548.101 (453) characters.[22] This content was coded for *Narrative Evaluation Favorability* as described in the "Dependent Variable" section of this paper.

Before formally testing my hypotheses, I examine the means of the conditions and nature of the data for *Narrative Evaluation Favorability*. Figure 4 contains a graphical representation of the means by condition for *Narrative Evaluation Favorability*. I tabulate the means, standard deviations, and sample size of the data in Table 3. The pattern of means appears to be consistent with my hypotheses. The mean of the positive directional goal conditions (mean = 4.92) is above the mean of the negative directional goal conditions (mean = 3.78). This pattern is consistent with theory that the supervisors' narrative evaluations were impacted by motivated reasoning. Also, the difference between the positive and negative directional goal conditions is smaller in the presence of rating (4.56 – 3.97 = 0.59) compared to the difference without ratings (5.23 - 3.52 = 1.71). This change due the presence of rating appears to be driven by a decrease in favorability

[21] In untabulated analyses, I find that inferences remain the same when dropping the five participants that answered this question incorrectly. All findings in Table 3, Table 4, and Table 5 replicate when dropping the five participants and the results become more statistically significant for some tests.

[22] To investigate whether the length of the narrative evaluations differed based on condition, I performed an untabulated ANOVA of the natural log of the number of words in the narrative on *Directional Goal*, *Rating Type*, and the interaction between the two. I find no differences in number of words by condition (all p's above .20). I perform a similar ANOVA investigating the effects of the manipulated variables on the natural log of the number of characters in the narratives and find no differences by condition (all p's above 0.15).

for the positive directional goal (5.23 to 4.56) and an increase in favorability for the negative directional goal (3.52 to 3.97). These descriptive results are directionally consistent with H1 and H2.

2.4.3 Hypothesis Tests

To formally test the hypotheses, I use a combination of Analysis of Variance (ANOVA) and follow-up simple effects tests. The results of the two-way ANOVA of *Narrative Evaluation Favorability* with *Directional Goal*, *Rating Type*, and the interaction as factors are reported in Table 3, Panel B. I find a significant main effect of *Directional Goal* on *Narrative Evaluation Favorability* ($F_{1,185}$=23.41, p<0.01) and a significant interaction between *Directional Goal* and *Rating Type* on *Narrative Evaluation Favorability* ($F_{1,185}$=4.32, p=0.039). This provides some evidence supporting H1 and H2. Follow-up simple effect tests provide additional insight into the directions of these effects. Without a rating, I find narrative evaluation favorability in the positive directional goal condition is higher than in the negative goal condition, on average (difference=1.609, t_{185}=4.957, two-tailed p<0.01). I also find a similar significant effect of *Directional Goal* when a rating is present, but the effect is much weaker (difference=0.642, t_{185}=1.926, two-tailed p=0.056). In combination, these effects support H1 and shed some initial light on H2. The reduced significance of the directional goal effect when a rating is present is consistent with theory that ratings reduce the effects of directional goals on the narrative evaluation. Since the impact of reasonableness constraints on directional goals depends on the direction of the goal, I also test the effects of ratings under positive and negative directional goals separately. The simple effect of including a rating when there is a positive directional goal is significantly negative (difference= -0.722, t_{185}= -2.214, two-tailed p=0.028). This provides evidence that the act of giving a rating imposes reasonableness constraints on the narrative evaluations which results in less favorable narrative evaluations under a positive directional

goal.[23] In other words, giving a rating reduces the effect of the directional goal on the favorability of the narrative evaluation. While not statistically significant, I find that the simple effect of including a rating with a negative directional goal fits the pattern predicted by my hypothesis (difference= 0.244, t_{185}= 0.737, one-tailed p=0.231). Under a negative directional goal, I expected *Narrative Evaluation Favorability* to be higher when a rating was required relative to no rating. This simple effect does not reach conventional levels of significance, but future research may investigate this effect further to understand whether the asymmetric magnitude of the effects of reasonableness constraints in this study is meaningful and what might cause the asymmetry.

2.4.4 Supplemental Analyses

2.4.4.1 Alternative Dependent Variable – Item Analysis

To better understand the effects of ratings and directional goals on narrative evaluations, I consider alternative dependent variables that divide the narrative evaluations into components rather than the single holistic measure, *Narrative Evaluation Favorability*. While the theory in this paper is discussed in broad terms about the overall favorability of the content, I expect the movement of overall favorability to be driven by the subcomponents of positive comments and negative comments in the narrative evaluations. If ratings restrict motivated reasoning in narrative evaluations, I expect the number of positive or negative comments in the narrative evaluation to be impacted by this restriction. For participants with positive (negative) directional goals, I expect those required to give a rating will provide fewer (more) positive comments and more (fewer) negative comments compared to those not required to give a rating.

[23] As additional evidence that participants narrative evaluations were constrained to matched the favorability of the ratings, I find that *Narrative Evaluation Favorability* and the ratings provided by participants in the *Rating* conditions are significantly correlated (Pearson Correlation = 0.742, $t_{df=90}$=10.512, two-tailed p<0.01).

I perform item analysis on the narrative evaluations by first dividing each evaluation into distinct statements (Smither and Walker 2004). To do this, I went through each narrative evaluation, while blind to experimental conditions, and divided the comments into a number of distinct statements. Distinct statements do not need to be complete sentences. Some sentences contain multiple statements (e.g., "Tom plays guitar and sings very well" is two statements – one about his guitar playing and another about his singing ability). Other statements are made of multiple sentences, yet constitute a single statement (e.g., "First of all, Tom seemed to be dressed too casually for the occasion. He's not dressed poorly, but he's dressing as if he's performing at your local pub - this is a professional Cruise service, where higher quality is expected. Maybe not a full suit, but I definitely would have liked to have seen Tom wearing khakis, and tuck in his shirt." – This is coded as a single statement). The mean (median) number of statements for the sample is 7.49 (7).

I coded each statement about Tom's performance as positive, negative, or neutral. I summed the number of positive statements for each evaluation to create *Positive Statements* and summed the number of negative statements for each evaluation to create *Negative Statements*.

In Table 4, I present descriptive statistics and the two-way ANOVA of *Positive Statements* with *Directional Goal*, *Rating Type*, and the interaction as factors. Consistent with the theory for H1, I find a significant main effect of *Directional Goal* on *Positive Statements* ($F_{1,185}$=20.01, p<0.01). Looking at the means in Table 4, we see that those participants with a positive directional goal included more positive statements in their narrative evaluation. However, I fail to find support for an interaction between *Directional Goal* and *Rating Type* as predicted by the theory in H2 ($F_{1,185}$=0.601, p=0.439). The simple effect of including a rating when there is a positive directional goal is significantly negative as theory would predict with ratings reducing the upward effect of the positive directional goal (difference= -0.707, t_{185}= -1.689, one-tailed p=0.047). However, the simple effect of including a rating for those with a positive directional goal is insignificant (difference= -0.244, t_{185}= -0.575, two-tailed p=0.566).

These results do not provide convincing evidence that the effect of directional goals on *Positive Statements* is reduced by the presence of a rating.

In Table 5, I present descriptive statistics and the two-way ANOVA of *Negative Statements* with *Directional Goal*, *Rating Type*, and the interaction as factors. As expected, I find a significant main effect of *Directional Goal* on *Negative Statements* with participants with a negative directional goal including more negative statements in their narrative evaluation, on average ($F_{1,185}$=20.17, p<0.01). I also find a marginally significant interaction between *Directional Goal* and *Rating Type*, suggesting that the effect of directional goals on the number of negative statements in the narrative evaluations depends on whether the supervisor also gives a rating ($F_{1,185}$=3.34, p=0.066). To investigate whether the direction of this interaction is consistent with H2, I follow-up the ANOVA with simple effects tests in Table 5, Panel C. Without a rating, having a positive directional goal results in a significant decrease in the number of negative comments (difference= -1.990, t_{185}= -4.546, two-tailed p<0.01). With a rating, the effect of having a positive directional goal is statistically significant but smaller in magnitude (difference= -0.828, t_{185}= -1.842, two-tailed p=0.067). Under a positive directional goal, the presence of a rating appears to increase the number of negative statements relative to no rating (difference= 0.618, t_{185}= 1.403, one-tailed p=0.081). This marginally significant effect is consistent with H2 that predicts supervisors who need to provide a rating will provide narrative evaluations that are less aligned with their directional goal. Under a negative directional goal, participants giving a rating provided fewer negative comments although the results don't reach conventional levels of significance (difference= -0.544, t_{185}= -1.217, one-tailed p=0.113). The overall pattern of results for the number of negative comments is consistent with the theory that narrative evaluations tend to be consistent with the supervisor's directional goal but the effect of directional goals on narrative evaluations decreases when the supervisor needs to provide a rating.

2.4.4.2 Analysis of Ratings

The theory development for H2 suggests that numerical ratings will be less affected by directional goals. Unfortunately, the large number of differences between the numerical scale and the coded narrative evaluation favorability makes direct comparisons of magnitude unwise. While I have no direct evidence that the numerical ratings were less affected by directional goals, I do test whether directionally goals had a measurable effect on the numerical ratings using an untabulated ordinal logistic regression[24] of numerical rating on *Directional Goal*. I use ordinal logistic regression because the labeled numerical rating scale does not meet the assumptions of an interval scale, however the order from 1 to 5 is meaningful with higher numbers reflecting more favorable evaluations. I fail to find a significant effect of *Directional Goal* on the numerical rating ($\beta = 0.529$, $t_{90} = 1.310$, two-tailed $p = 0.193$). The lack of a detectable effect of *Directional Goal* on numerical ratings when the same manipulation significantly affected the narrative evaluations provides some interesting although inconclusive evidence that the numerical ratings were more resistant to the effects of directional goals.

2.4.4.3 Effects of Narratives on Ratings

The bulk of this paper has discussed the anticipated impact of ratings on narrative evaluations. However, one might also be interested in the effects of narrative evaluations on ratings. In the primary experiment, I included two additional conditions where participants provided a rating, but did not provide a narrative evaluation. These additional 93 participants from the same participant pool split across negative and positive conditions are a benchmark to compare to the conditions where participants responded to both *Rating* and *Narrative* prompts. In untabulated analyses using an ordinal logistic regression of the supervisor's rating of the employee on *Directional Goal*, whether or not the supervisor provides a narrative, and the interaction between these two variables, I find no evidence that either of these manipulations or

[24] I use the *MASS* package for R to run the ordinal logistic regression.

the interaction affected the performance rating (all p's>0.40). While the purpose of this study was not to investigate the effects of narratives on numerical ratings, I note that no effects were detected. Since the study was not designed specifically to test the effects of narrative on ratings, these initial findings should not preclude future research from investigating these effects.

2.4.4.4 Perceived Similarities and Differences between Ratings and Narratives

In my study, I asked participants in the rating plus narrative evaluation conditions to answer additional questions about how they perceived ratings and narrative evaluations. Participants responded to the following two questions on 11-point scales (-5 Completely Disagree, 0 Neither Agree Nor Disagree, 5 Completely Agree): (1) "I felt the need to make my rating and written evaluation consistent with each other" and (2) "My rating and written evaluation were different from each other". Participants reported they tended to agree that they needed to make their rating and written evaluations consistent with each other with responses being significantly higher than the midpoint of the scale (mean = 3.130, H_0: $\mu=0$, $t_{91}=14.882$, two-tailed $p<.01$). Similarly, participants reported they tended to disagree with the statement that their rating and written evaluations were different from each other with responses being significantly lower than the midpoint of the scale (mean = -2.207, H_0: $\mu=0$, $t_{91}= -7.691$, two-tailed $p<.01$). These results provide evidence that participants view ratings and narrative as telling a similar story about employee performance (i.e., they match).

Since the primary goal of this study was to understand the effects of ratings on narrative evaluations, I intentionally presented rating scale at the top of the page with the narrative evaluation below it. I asked participants, "Which of the two pieces of the evaluation did you fill out first?", on an 11-point scale (0 Written Evaluation, 5 I filled them out simultaneously, 10 Numerical Rating). As expected, given the ordering of the two tasks, on average, participants suggested they filled out the numerical rating before the written evaluation (mean = 7.065, H_0: $\mu=5$, $t_{91}= 4.899$, two-tailed $p<.01$). To capture their perceptions of importance of the two evaluations, they also answer the following question on an 11-point scale (0 Written Evaluation is

clearly more important, 5 Equally important, 10 Numerical Rating is clearly more important): "Which of the two pieces of the evaluation did you think was more important?" They perceive written evaluations to be more important than numerical ratings (mean = 1.880, H_0: $\mu=5$, $t_{91}= -14.242$, two-tailed $p<.01$). Last, they responded to the following question on an 11-point scale (0 - 100% Written Evaluation Effort, 5 -Equal Effort, 10 - 100% Numerical Rating Effort): "Which of the two pieces of the evaluation did you put more thought and effort into?". Consistent with perceptions of relative importance, participants also suggest they put more thought and effort into the narrative evaluation than the numerical rating (mean = 1.924, H_0: $\mu=5$, $t_{91}= -14.28$, two-tailed $p<.01$). While these supplementary questions are not conclusive evidence, they suggest that evaluators attempted to match their ratings and narratives, filled out their ratings before the narratives, and perceived narratives to be important. These results taken together with previously reported findings in this paper tell an intriguing story of unintended consequences. Specifically, participants may be surprised to learn that the substantially less important and low effort rating is influencing narratives in a significant way. Similarly, from the opposite point of view, participants may be surprised by the strong effects of directional goals on their narrative evaluations and the ability of a simple rating judgment to reduce this effect.

2.5 Conclusions

Understanding the effects of ratings on narrative evaluations is particularly important given the push to remove subjective performance ratings in practice. Using an experiment, I investigate the effects of ratings on narrative evaluations using a laboratory experiment. I find that the requirement to provide ratings does affect narrative evaluations of employee performance, specifically the favorability of those narrative evaluations. While I do find narrative evaluations are influenced by directional goals, I also find that requiring an evaluator to provide a numerical performance rating reduces the effect of directional goals on narrative evaluation favorability. These results are consistent with motivated reasoning theory. Specifically, narrative

evaluation favorability is affected by directional goals, but this effect is reduced in the presence of reasonableness constraints (i.e., the rating process).

While firms are getting rid of performance evaluations, they may be failing to consider a benefit to retaining these performance ratings in their performance management system. The tendency for preference-consistent behavior by supervisors in free-form narrative evaluations appears to be mitigated by performance ratings.

This study suggests interesting areas for future research. In this study, I focus solely on one dimension of narrative evaluations, favorability. Given the rich nature of narrative evaluations, future research could investigate similar questions using different dimensions of narrative evaluation (Doyle 2018). Also, this study establishes narrative performance evaluations as an important outcome variable in the management accounting literature. Future research could investigate how other control system functions impact subjective narrative evaluations.

In this study, I selected a specific rating scale and set of labels to be a reasonable representation of ratings used in practice. However, the exact rating scale and labels used in firms varies depending on the firm's needs. Rating scales in practice could include more or fewer categories, fewer labels, different labels, among other differences from the scale used in my study. Future research can investigate whether specific types of scales and labels affect narrative evaluations differently. Some firms even avoid the use of numbers but maintain the same labels. I argue that these types of rating scales are very similar to numbered scales because it requires a summary judgment that group employees into an ordinal category with a specified meaning. However, I did not separately manipulate the presence of numbers on the scale. Future research can dissect the parts of the scale and investigate whether the effects are driven by a specific piece of the rating scale.

2.6 Appendix – Example Narrative Evaluations from Participants

Example #1 – Negative Goal Condition Example

Narrative Evaluation Favorability – 2

Positive Statements – 2

Negative Statements – 7

"Overall, I feel like Tom is a good performer, but there were definitely a few glaring issues. First of all, Tom seemed to be dressed too casually for the occasion. He's not dressed poorly, but he's dressing as if he's performing at your local pub - this is a professional Cruise service, where higher quality is expected. Maybe not a full suit, but I definitely would have liked to have seen Tom wearing khakis, and tuck in his shirt. Also, Tom just seemed a bit bored during his performance. It is a sadder toned song, but I would have liked to have seen a bit more emotion. He was constantly either looking down at his guitar/feet, or he had a blank stare on his face that showed no emotion. Lastly, I feel like Tom sang the song a little too quickly. I know it's a cover, but I would have liked to have seen him take it slower to showcase the emotion of the song and the power of his voice."

Example #2 – Positive Goal Condition Example

Narrative Evaluation Favorability – 5.5

Positive Statements – 5

Negative Statements – 3

"I felt as though Tom played a recent and nice song choice. He has a great voice and plays the guitar very well. Although it was a good performance, it seemed as though he had to look down at the lyrics or guitar chords a lot which interfered with Tom walking around and getting more involved with the audience. Also, Tom's beautiful raspy voice would be well accompanied by a smile as well as some more emotion and feeling in some of the words. Overall, Tom did a fine job but could most definitely be outstanding after some tune ups."

2.7 Figures and Tables for Chapter 2

FIGURE 4 - Graph of Cell Means for Narrative Evaluation Favorability from Table 3

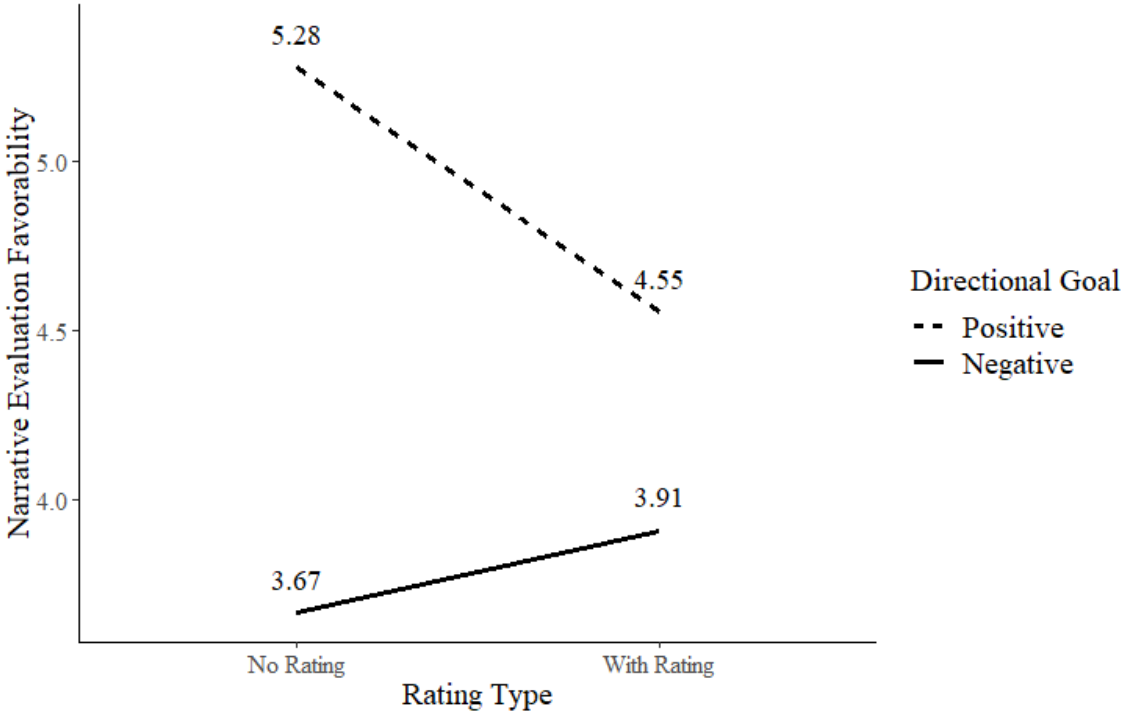


Figure Note – This figure graphically displays the effect of Directional Goal and Rating Type on supervisors' narrative evaluation favorability. The means are described with formal hypothesis tests in Table 3. All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Narrative Evaluation Favorability*, is a measure of the favorability of each supervisor's narrative performance evaluation as coded by two independent coders on a 1 to 7 scale. *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*with rating*) of employee performance while other did not provide a rating (*no rating*). *Directional Goal* was manipulated as *Positive* or *Negative* using advice from the participants' hypothetical boss that recommended being either more nice or more harsh in their evaluations, respectively.

TABLE 3 - Analysis of Narrative Evaluation Favorability

Panel A: *Narrative Evaluation Favorability* Mean (Standard Deviation) [Sample Size]

	Rating Type		
Directional Goal	No Rating	Rating	Overall
Positive	5.28	4.55	4.92
	(1.34)	(1.67)	(1.55)
	[49]	[47]	[96]
Negative	3.67	3.91	3.78
	(1.68)	(1.68)	(1.68)
	[48]	[45]	[93]
Overall	4.48	4.24	4.36
	(1.72)	(1.70)	(1.71)
	[97]	[92]	[189]

Panel B: Two-way ANOVA Model of *Narrative Evaluation Favorability*

Source of Variation	SS	df	F	p-value
Intercept	3575.5	1	1400.037	<.01
Directional Goal	59.8	1	23.412	<.01
Rating Type	2.7	1	1.055	0.306
Directional Goal x *Rating Type*	11.0	1	4.319	0.039
Error	472.5	185		

Panel C: Simple Effects

Simple Effect	Estimate	$t_{df=185}$	p
Effect of Directional Goal with No Rating	1.609	4.957	<.01
Effect of Directional Goal with Rating	0.642	1.926	0.056
Effect of Rating with Positive Directional Goal	-0.722	-2.214	0.028
Effect of Rating with Negative Directional Goal	0.244	0.737	0.462

Table Notes - All p-values listed in the table are two-tailed. This table presents descriptive statistics, an ANOVA model, and simple effects for *Narrative Evaluation Favorability*. All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Narrative Evaluation Favorability*, is a measure of the favorability of each supervisor's narrative performance evaluation as coded by two independent coders on a 1 to 7 scale. *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*with rating*) of employee performance while other did not provide a rating (*no rating*). *Directional Goal* was manipulated as *Positive* or *Negative* using advice from the participants' hypothetical boss that recommended being either more nice or more harsh in their evaluations, respectively.

TABLE 4 - Analysis of Number of Positive Statements in Narrative Evaluations

Panel A: *Positive Statements* Mean (Standard Deviation) [Sample Size]

	Rating Type		
Directional Goal	No Rating	Rating	Overall
Positive	4.90	4.19	4.55
	(1.88)	(1.90)	(1.91)
	[49]	[47]	[96]
Negative	3.33	3.09	3.22
	(2.36)	(2.01)	(2.19)
	[48]	[45]	[93]
Overall	4.12	3.65	3.89
	(2.27)	(2.02)	(2.16)
	[97]	[92]	[189]

Panel B: Two-way ANOVA Model of *Positive Statements*

Source of Variation	SS	df	F	p-value
Intercept	2839.4	1	676.852	<.01
Direction Goal	83.95	1	20.012	<.01
Rating Type	10.67	1	2.544	0.112
Directional Goal x *Rating Type*	2.52	1	0.601	0.439
Error	776.08	185		

Panel C: Simple Effects

Simple Effect	Estimate	$t_{df=185}$	p
Effect of Directional Goal with No Rating	1.565	3.762	<.01
Effect of Directional Goal with Rating	1.103	2.581	0.011
Effect of Rating with Positive Directional Goal	-0.707	-1.689	0.093
Effect of Rating with Negative Directional Goal	-0.244	-0.575	0.566

Table Notes - All p-values listed in the table are two-tailed. This table presents descriptive statistics, an ANOVA model, and simple effects for *Positive Statements*. All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Positive Statements*, is a count of the number of positive statements in each narrative evaluation as coded by the author. *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*with rating*) of employee performance while other did not provide a rating (*no rating*). *Directional Goal* was manipulated as *Positive* or *Negative* using advice from the participants' hypothetical boss that recommended being either more nice or more harsh in their evaluations, respectively.

TABLE 5 - Analysis of Number of Negative Statements in Narrative Evaluations

Panel A: *Negative Statements* Mean (Standard Deviation) [Sample Size]

	Rating Type		
Directional Goal	No Rating	Rating	Overall
Positive	2.51	3.13	2.81
	(1.72)	(2.46)	(2.13)
	[49]	[47]	[96]
Negative	4.50	3.96	4.24
	(2.21)	(2.17)	(2.20)
	[48]	[45]	[93]
Overall	3.49	3.53	3.51
	(2.21)	(2.35)	(2.27)
	[97]	[92]	[189]

Panel B: Two-way ANOVA Model of *Negative Statements*

Source of Variation	SS	df	F	p-value
Intercept	2343.92	1	504.573	<.01
Directional Goal	93.69	1	20.169	<.01
Rating Type	0.06	1	0.014	0.907
Directional Goal x *Rating Type*	15.93	1	3.430	0.066
Error	859.39	185		

Panel C: Simple Effects

Simple Effect	Estimate	$t_{df=185}$	p
Effect of Directional Goal with No Rating	-1.990	-4.546	<.01
Effect of Directional Goal with Rating	-0.828	-1.842	0.067
Effect of Rating with Positive Directional Goal	0.618	1.403	0.162
Effect of Rating with Negative Directional Goal	-0.544	-1.217	0.225

Table Notes - All p-values listed in the table are two-tailed. This table presents descriptive statistics, an ANOVA model, and simple effects for *Negative Statements*. All participants (acting as supervisors) provided a narrative evaluation of the employee. The dependent variable, *Negative Statements*, is a count of the number of negative statements in each narrative evaluation as coded by the author. *Rating Type* was manipulated by requiring some participants to provide a numerical performance rating (*with rating*) of employee performance while other did not provide a rating (*no rating*). *Directional Goal* was manipulated as *Positive* or *Negative* using advice from the participants' hypothetical boss that recommended being either more nice or more harsh in their evaluations, respectively.

BIBLIOGRAPHY

Adler, S., M. Campion, A. Colquitt, A. Grubb, K. Murphy, R. Ollander-Krane, and E. D. Pulakos. 2016. Getting Rid of Performance Ratings: Genius or Folly? A Debate. *Industrial and Organizational Psychology* 9 (02): 219–252.

Amer, T., K. Hackenbrack, and M. Nelson. 1994. Between-Auditor Differences in the Interpretation of Probability Phrases. *Auditing: A Journal of Practice and Theory* 13 (1): 126-136.

Amer, T., K. Hackenbrack, and M. Nelson. 1995. Context-Dependence of Auditors' Interpretations of the SFAS No. 5 Probability Expressions. *Contemporary Accounting Research* 12 (1): 25-39.

Asay, H. S., R. Libby, and K. Rennekamp. 2018. Firm performance, reporting goals, and language choices in narrative disclosures. *Journal of Accounting and Economics* 65 (2): 380–398.

Bentley, J. W. 2019. Decreasing Operational Distortion and Surrogation Through Narrative Reporting. *Accounting Review* 94 (3): 27-55.

Bol, J. C. 2008. Subjectivity in Compensation Contracting. *Journal of Accounting Literature; Gainesville* 27: 1–24.

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

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

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

Bradshaw, M. T., L. F. Lee, and K. Peterson. 2016. The Interactive Role of Difficulty and Incentives in Explaining the Annual Earnings Forecast Walkdown. *Accounting Review* 91 (4): 995-1021.

Brutus, S. 2010. Words Versus Numbers: A Theoretical Exploration of Giving and Receiving Narrative Comments in Performance Appraisal. *Human Resource Management Review* 20 (2010): 144-157.

Buckingham, M., and A. Goodall. 2015. Reinventing Performance Management: How One Company is Rethinking Peer Feedback and the Annual Review, and Trying to Design a System to Fuel Improvement. *Harvard Business Review* (4): 40-45.

Cappelli, P., and A. Tavis. 2016. The performance management revolution: the focus is shifting from accountability to learning. *Harvard Business Review* (10).

Casas-Arce, P., S. M. Lourenço, and F. A. Martínez-Jerez. 2017. The Performance Effect of Feedback Frequency and Detail: Evidence from a Field Experiment in Customer

Satisfaction: The Performance Effect of Feedback Frequency and Detail. *Journal of Accounting Research* 55 (5): 1051–1088.

Chen, Y., J. Jermias, and T. Panggabean. 2016. The Role of Visual Attention in the Managerial Judgment of Balanced-Scorecard Performance Evaluation: Insights from using an eye-tracking device. *Journal of Accounting Research* 54 (1): 113-145.

Cicchetti, D. V. 1994. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*. 6 (4): 284-290.

Corporate Executive Board (CEB). 2016. The real impact of eliminating performance ratings. Corporate Executive Board (CEB Global). Available at: https://emtemp.gcom.cloud/ngw/globalassets/en/human-resources/documents/trends/eliminate-performance-ratings.pdf

Cuccia, A., K. Hackenbrack, and M. Nelson. 1995. The Ability of Professional Standards to Mitigate Aggressive Reporting. *Accounting Review* 70 (2): 227–248.

Daft, R. and J. Wiginton. 1979. Language and Organization. *Academy of Management Review* 4 (2): 179–191.

David, E. 2013. Examining the Role of Narrative Performance Appraisal Comments on Performance. *Human Performance* 26 (5): 430–450.

Decotiis, T., and A. Petit. 1978. The Performance Appraisal Process: A Model and Some Testable Propositions. *The Academy of Management Review* (3): 635.

Demski, J. S., and G. A. Feltham. 1976. *Cost determination: a conceptual approach.* Ames, IA: Iowa State University Press.

Edmonds, T., C. Edmonds, M. Edmonds, J. Edmonds, and P. Olds. 2020. *Fundamental Managerial Accounting Concepts*. New York: McGraw-Hill Education.

Erdogan, B. 2002. Antecedents and consequences of justice perceptions in performance appraisals. *Human Resource Management Review* 12 (4): 555–578.

Gorman, C. A., J. P. Meriac, S. G. Roch, J. L. Ray, and J. S. Gamble. 2017. An exploratory study of current performance management practices: Human resource executives' perspectives. *International Journal of Selection and Assessment* 25: 193-202.

Hackenbrack, K., and M. Nelson. 1996. Auditors' Incentives and Their Application of Financial Accounting Standards. *Accounting Review* 71 (1): 43-59.

Hales, J. 2007. Directional Preferences, Information Processing, and Investors' Forecasts of Earnings. *Journal of Accounting Research* 45 (3): 607-628.

Hallgren, K. A. 2012. Computing Inter-Rater Reliability for Observational Data: an overview and tutorial. *Tutor Quant Methods Psychol* 8 (1): 23–34.

Hannan, R. L., R. Krishnan, and A. H. Newman. 2008. The Effects of Disseminating Relative Performance Feedback in Tournament and Individual Performance Compensation Plans. *Accounting Review* 83 (4): 893–913.

Hannan, R. L., G. P. McPhee, A. H. Newman, and I. D. Tafkov. 2013. The Effect of Relative Performance Information on Performance and Effort Allocation in a Multi-Task Environment. *Accounting Review* 88 (2): 553–575.

Hornsey, M. J., and A. Imani. 2004. Criticizing Groups from the Inside and the Outside: An Identity Perspective on the Intergroup Sensitivity Effect. *Personality and Social Psychology Bulletin* 30 (3): 365–383.

Hornsey, M. J., T. Oppes, and A. Svensson. 2002. "It's OK if we say it, but you can't": responses to intergroup and intragroup criticism. *European Journal of Social Psychology* 32 (3): 293–307.

Ittner, C. D., D. F. Larcker, and M. W. Meyer. 2003. Subjectivity and the Weighting of Performance Measures: Evidence from a balanced scorecard. *Accounting Review* 78 (3): 725–758.

Jawahar, I. M., and C. R. Williams. 1997. Where All the Children Are Above Average: The Performance Appraisal Purpose Effect. *Personnel Psychology* 50 (4): 905–926.

Kluger, A. N., and A. DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119 (2): 254-284.

Kunda, Z. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108 (3): 480-498.

Lerner, J. S., and P. E. Tetlock. 1999. Accounting for the effects of accountability. *Psychological Bulletin* 125 (2): 255–275.

Li, F. 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics* 45 (2–3): 221–247.

Libby, R., R. Bloomfield, and M. W. Nelson. 2002. Experimental research in financial accounting. *Accounting, Organizations and Society* 27 (8): 775–810.

Loftus, S., and L. J. Tanlu. 2018. Because of "Because": Examining the Use of Causal Language in Relative Performance Feedback. *Accounting Review* 93 (2): 277–297.

London, M. 2003. *Job Feedback: Giving, Seeking, and Using Feedback for Performance Improvement*. Mahwah, NJ: Lawrence Erlbaum Associates.

Merritt, A. C., D. A. Effron, S. Fein, K. K. Savitsky, D. M. Tuller, and B. Monin. 2012. The strategic pursuit of moral credentials. *Journal of Experimental Social Psychology* 48 (3): 774–777.

Miller, D. T., and D. A. Effron. 2010. Chapter Three - Psychological License: When it is Needed and How it Functions. In *Advances in Experimental Social Psychology*, edited by M. P. Zanna and J. M. Olson, 43:115–155. Academic Press.

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

Murphy, K. R., and J. N. Cleveland. 1995. *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage.

Peecher, M. E., M. D. Piercey, J. S. Rich, and R. M. Tubbs. 2010. The Effects of a Supervisor's Active Intervention in Subordinates' Judgments, Directional Goals, and Perceived Technical Knowledge Advantage on Audit Team Judgments. *Accounting Review* 85 (5): 1763-1786.

Piercey, M. D. 2009. Motivated Reasoning and Verbal vs. Numerical Probability Assessment: Evidence from an accounting context. *Organizational Behavior and Human Decision Processes* 108 (2009): 330-341.

Spilker, B., R. Worsham, and D. Prawitt. 1999. Tax Professionals' Interpretations of Ambiguity in Compliance and Planning Decision Contexts. *Journal of the American Taxation Association* 21 (2): 75-89.

Rock, D., J. Davis, and B. Jones. 2014. Kill Your Performance Ratings. *strategy+business* 76.

Rock, D., and B. Jones. 2015. Why More and More Companies Are Ditching Performance Ratings. *Harvard Business Review*, September 8.

Siegel-Jacobs, K., and J. F. Yates. 1996. Effects of Procedural and Outcome Accountability on Judgment Quality. *Organizational Behavior and Human Decision Processes* 65 (1): 1–17.

Silverman, R. E. 2016. Management: GE Scraps Staff Ratings to Spur Feedback. *Wall Street Journal, Eastern edition; New York, N.Y.*, July 27.

Smither, J. W., and A. G. Walker. 2004. Are the Characteristics of Narrative Comments Related to Improvement in Multirater Feedback Ratings Over Time? *Journal of Applied Psychology* 89 (3): 575–591.

Speer, A. B. 2018. Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology* 71 (3): 299–333.

Swieringa, R. J., and K. E. Weick. 1981. Interfaces between Management Accounting and Organizational Behavior. *Exchange: The Organizational Behavior Teaching Journal* VI (3): 25-33.

Tan, H., and K. Jamal. 2001. Do Auditors Objectively Evaluate Their Subordinates' Work? *Accounting Review* 76 (1): 99-110.

Tayler, W. B. 2010. The Balanced Scorecard as a Strategy-Evaluation Tool: the effects of implementation involvement and a causal-chain focus. *Accounting Review* 85 (3): 1095–1117.

White, R. M. 2019. *There and Back Again – The performance evaluation effects of going to and returning from part-time status*. Working Paper, Arizona State University.

Wilks, T. J. 2002. Predecisional Distortion of Evidence as a Consequence of Real-Time Audit Review. *Accounting Review* 77 (1): 51–71.

Wilson, K. Y. 2010. An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations* 63 (12): 1903–1933.

Woods, A. 2012. Subjective Adjustments to Objective Performance Measures: the influence of prior performance. *Accounting, Organizations and Society* 37 (2012): 40–425.

Zedeck, S., and W. F. Cascio. 1982. Performance Appraisal Decisions as a Function of Rater Training and Purpose of the Appraisal. *Journal of Applied Psychology* 67 (6): 752–758.