Calibration Committees and Rating Distribution Guidance Effects on Leniency Bias in Subjective Performance Evaluations

by

Katharine Elizabeth Patterson

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Accounting

Waterloo, Ontario, Canada, 2023

EXAMING COMMITTEE MEMBERSHIP

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Johnny Jermias Professor of Accounting
Supervisor	Krista Fiolleau Associate Professor of Accounting
Internal Member	Bradley Pomeroy Associate Professor of Accounting
Internal Member	Adam Vitalis Assistant Professor of Accounting
Internal-External Member	Abigail Scholer Professor of Psychology

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

ABSTRACT

Firms use both calibration committees and rating distribution guidance to reduce leniency bias in subjective performance ratings. Leniency bias is the tendency to provide subordinates with higher ratings than deserved which can weaken the link between incentives and effort, leading to suboptimal and subordinate performance. I employ a 2x2 online experiment to assess how the presence versus absence of peer calibration committees [PCCs] and rating distribution guidance [RDG] affects leniency bias present in supervisors' ratings of subordinates' performance. I find support that supervisors may display more leniency in ratings prepared in anticipation of a PCC, especially among low performers. As the increased bias appears to impact low-performers, this may create additional fairness concerns for moderate and high-performers, which could demotivate these subordinates. Next, I find support that rating distribution guidance does have a main effect of reducing the leniency bias displayed among low and high performers. Further, using planned contrast testing, I find support for my predicted pattern of results for low performers. That is, the presence of a PCC has a main effect of increasing leniency bias, the presence of RDG has the main effect of reducing leniency bias, and the interactive effect such that when a PCC is present, the presence of RDG weakens the effect of PCCs on leniency bias. This finding indicates that rating distribution guidance may be helpful in settings with a PCC.

Keywords: Subjective performance evaluation; leniency bias; calibration committees; rating distribution guidance; social comparison; impression management; injunctive norm

ACKNOWLEDGEMENTS

I am grateful to all those who provided feedback, guidance, and support throughout this process. I would like to offer particular thanks to Krista Fiolleau for her invaluable support, it is deeply appreciated. Thank you to the other members of my examining committee, Johnny Jermias, Brad Pomeroy, Abigail Scholer, and Adam Vitalis for their time, expertise and insightful feedback. I also offer particular thanks to Alan Webb and Adam Presslee, as well as, the students and faculty at the University of Waterloo and the University of Vienna Economics and Business. I am also grateful to the CPA Ontario Centre for Performance Management and Education (CPMRE) at the University of Waterloo for providing funding for this project.

TABLE OF CONTENTS

LIST OF FIGURES......viii

LIST OF TABLES......ix

CHAPTER 1 : INTRODUCTION......1

CHAPTER 2 : LITERATURE REVIEW......9

2.1 INTRODUCTION......9
2.2 SUBJECTIVE PERFORMANCE EVALUATIONS......9
2.2.1 Subjective Performance Evaluations......9
2.3 LENIENCY BIAS IN SUBJECTIVE PERFORMANCE EVALUATIONS......11
2.3.1 Leniency Bias in Subjective Performance Evaluations......12
2.3.2 The Benefits and Costs of Leniency Bias......15
2.4 SOCIAL COMPARISON AND IMPRESSION MANAGEMENT IN PERFORMANCE EVALUATIONS......17
2.4.1 Social Comparison and Impression Management......17
2.5 INJUNCTIVE NORMS IN THE SUBJECTIVE EVALUATION PROCESS......20
2.6 CALIBRATION COMMITTEES......21
2.6.1 Calibration Committees......22
2.7 RATING DISTRIBUTION GUIDANCE (RDG)......24
2.7.1 Rating Distribution Guidance......24
2.8 RATING DISTRIBUTION GUIDANCE IN ENVIRONMENTS WHERE PEER CALIBRATION COMMITTEES ARE PRESENT......28
2.9 CONCLUSION......29

CHAPTER 3 : DEVELOPMENT OF HYPOTHESES......31

3.1 INTRODUCTION......31
3.2 THE EFFECTS OF A PEER CALIBRATION COMMITTEE ON LENIENCY BIAS......31
3.3 THE EFFECT OF RATING DISTRIBUTION GUIDANCE......35
3.4 THE INTERACTIVE EFFECT OF PEER CALIBRATION COMMITTEES AND RATING DISTRIBUTION GUIDANCE......37

CHAPTER 4 : RESEARCH METHOD......42

4.1 INTRODUCTION......42
4.2 EXPERIMENTAL DESIGN OVERVIEW......42
4.3 PARTICIPANTS......44
4.4 GENERAL INSTRUCTIONS AND SCENARIOS......46
4.4.1 General Instructions......46
4.4.2 Base Scenario......46
4.4.3 PCC Present - Scenario......48
4.4.4 RDG Present - Scenario......49
4.5 SUBORDINATE PROFILES......51
4.5.1 Subordinate Profile Presentation......51
4.5.2 The Number of Subordinate Profiles......53
4.5.3 The Performance Level Distribution of Subordinate Profiles......54
4.6 DEPENDENT VARIABLES......57
4.6.1 Creating a Leniency Bias Measure......57
4.6.2 Process Measures......59
4.7 CONCLUSION......61

CHAPTER 5 : RESULTS......62

5.1 INTRODUCTION......62
5.2 EXPERIMENTAL PARTICIPANTS' DEMOGRAPHIC DATA AND CONTROL VARIABLES......62
5.2.1 Participants Sample......62

5.2.2 Demographic Information 63
5.2.3 Leniency Bias 64
5.2.4 Correlation Analysis 64
5.3 TEST OF HYPOTHESES 64
5.3.1 Method of Analysis 64
5.3.2 Hypothesis One 68
5.3.3 Hypothesis Two 69
5.3.4 Hypothesis Three 69
5.4 ADDITIONAL TESTING – THEORETICAL MECHANISMS 70
5.4.1 Principle Components Analysis 70
5.5 SUMMARY 72
CHAPTER 6 : CONCLUSION 74
6.1 INTRODUCTION 74
6.2 DISCUSSION OF HYPOTHESES TESTING RESULTS 74
6.3 LIMITATIONS AND OPPORTUNITIES FOR FUTURE RESEARCH 75
6.4 CONCLUSIONS 76
REFERENCES 78
FIGURES 96
TABLES 100
APPENDICES 115
APPENDIX A - EXPERIMENTAL DESIGN OVERVIEW 115
APPENDIX B - GENERAL INSTRUCTIONS 116
APPENDIX C – SCENARIOS 117
APPENDIX D - SUBORDINATE PROFILES 119
APPENDIX E - POST EXPERIMENTAL AND DEMOGRAPHIC QUESTIONS 131
APPENDIX F - PILOT TEST ONE 134
APPENDIX G - PILOT TEST TWO 141
APPENDIX H – SUBORDINATE MIDPOINT RATINGS 150
APPENDIX I - VARIABLE NAMES AND DEFINITIONS 151

LIST OF FIGURES

FIGURES..78

Figure 1 - Hypothesis One 96
Figure 2 - Hypothesis Two...97
Figure 3 - Hypothesized Conceptual Model 98
Figure 4 - Graphical Depiction of Hypotheses 99

LIST OF TABLES

TABLES 100

TABLE 1 – DESCRIPTIVE STATISTICS 100
TABLE 2 - CORRELATIONS 105
TABLE 3 - HYPOTHESIS TEST.......... 106
TABLE 4 - HYPOTHESIS TESTING – SUBORDINATES PARTITIONED BY PERFORMANCE LEVEL 107
TABLE 5 - ADDITIONAL HYPOTHESIS TESTING – PLANNED CONTRAST TESTING – LOW PERFORMERS 108
TABLE 6 - ADDITIONAL HYPOTHESIS TESTING – PLANNED CONTRAST TESTING – MODERATE PERFORMERS 110
TABLE 7 - ADDITIONAL HYPOTHESIS TESTING – PLANNED CONTRAST TESTING – HIGH PERFORMERS.......... 111
TABLE 8 – SUMMARY STATISTICS – SOCIAL COMPARISON, IMPRESSION MANAGEMENT AND NORMS 112
TABLE 9 – FACTOR PRINCIPAL COMPONENTS ANALYSIS.......... 113
TABLE 11 – TESTING OF THEORETICAL CONSTRUCTS.......... 114

CHAPTER 1: INTRODUCTION

Performance evaluations are a crucial component of an organization's performance management systems. Evaluations provide subordinates with performance feedback information and help inform the promotion, retention, and compensation decisions made by supervisors (Bol, 2011; Mercer, 2013; Moers, 2005).[1] This study examines two controls used by organizations to help reduce bias in the performance evaluation process. Specifically, I investigate the effects of peer calibration committees [PCCs] and rating distribution guidance [RDG] on supervisors' leniency bias in subjective performance evaluation.[2]

Using subjectivity is often essential to evaluate a subordinate's contributions to the firm. Subjectivity can allow supervisors to provide more accurate, informative, and timely evaluations (Gibbs et al., 2004, p. 411). For example, subjectivity is useful when evaluating professional employees when it may not be possible or practical to evaluate the quantity or quality of their work objectively, and thus subjectivity is needed (e.g., auditors, consultants, managers, or information technology project staff; Bol, 2008; Gibbs et al., 2004). Performance evaluations may incorporate subjectivity in various ways; however, this paper will focus on the use of subjective judgements to evaluate subordinate performance.

Despite subjectivity being useful in an evaluation of subordinates' performances, subjectivity may also introduce bias, thereby decreasing the accuracy of performance ratings (Bol, 2008, 2011; Gibbs et al., 2004; e.g., Prendergast & Topel, 1993). Biases in ratings can decrease employee effort, increase employee turnover, and impair employee performance; as they provide

[1] 'Supervisor' is used when referring to the manager evaluating the subordinates and 'subordinate' when referring to the subordinate the supervisor is evaluating.

[2] Leniency bias is the tendency to provide subordinates with higher ratings than they deserve based on their 'true' performance (Saal & Landy, 1977).

inaccurate feedback, weaken communication of the organization's goals, and cause some subordinates to perceive the evaluation process as unfair (Baker et al., 1988; Bol, 2008; Jawahar & Williams, 1997; Prendergast, 1999).[3] Management research firm CEB finds that ninety percent of both supervisors and human resource leaders are dissatisfied with their organization's annual performance evaluations and believe the process does not yield accurate information (Wilkie, 2015). As such, many practitioners and researchers focus on identifying controls that may help to reduce bias in performance evaluations. This paper examines one specific bias, leniency bias, commonly exhibited by supervisors when they use subjective judgements to evaluate their subordinates' performances (e.g., Bol, 2011; Moers, 2005). Specifically, I examine two controls used by organizations to minimize leniency bias in subjective performance evaluations; peer calibration committees [PCCs] and rating distribution guidance [RDG].

My first prediction examines the use of calibration committees. A calibration committee is a group of supervisors whose task is to collectively discuss and possibly adjust subordinate performance ratings. These committees are typically formed with the goal of reducing the bias that can occur when supervisors have final authority over performance ratings (Bol et al., 2019; Demeré et al., 2019; Grabner et al., 2020). Proponents of calibration committees, including many consultants, endorse them as a best practice to reduce bias and survey evidence shows that over 56% of organizations use calibration committees (Albert, 2017; Hastings, 2011; Mercer, 2013; Risher, 2014). While three main types of calibration committees exist in practice, this study will focus on peer-level calibration committees [PCCs], a committee comprised of peer supervisors (Bol et al., 2019).

[3] Biased ratings are inaccurate ratings resulting from systematic deviation from the true value of the measured variable (Merchant & Van der Stede, 2017).

Prior research presents some insights regarding the use of a PCC to reduce leniency bias. For example, Bol et al.’s (2019) field study finds that, on average, the performance ratings of subordinates are lower (i.e., have less leniency bias) *after* discussion by the PCC compared to *before* discussion by the PCC (Bol et al., 2019). However, while this evidence suggests PCCs may reduce the bias in ratings pre-to-post-committee review, it is not yet known how supervisors’ anticipation of a PCC will impact the extent of leniency bias in ratings prepared for the PCC’s consideration. This is also a critical factor to examine since the quality of the inputs to the PCC limits the quality of any output. Therefore, understanding the impact of a PCC on the ratings prepared by supervisors in anticipation of a committee’s review is essential to understanding a calibration committee’s impact on leniency bias.

A PCC creates a forum, or audience, for social comparison of managerial ability among committee members. As such, social comparison theory suggests that supervisors who anticipate reporting their ratings to a committee of their peers will seek to achieve a positive self-image (Beach & Tesser, 1995; D. Brown et al., 2007; Festinger, 1954). Consequently, by increasing social comparison pressures, a PCC will shift a supervisor’s focus more towards impression management (Bol et al., 2019; Grabner et al., 2020). Prior research finds that supervisors may seek to manage impressions by providing lenient ratings to their subordinates in an effort to signal their managerial ability to others, including their peers and higher-level managers (Bol et al., 2019; Ilgen et al., 1981; Merchant & Van der Stede, 2017; Rosaz & Villeval, 2012). Thus, due to increased social comparison and impression management desires, my first hypothesis predicts that a supervisor’s anticipation of reporting to a PCC will increase leniency bias compared to a supervisor who has final authority over the ratings.

My second prediction examines the impact of Rating Distribution Guidance [RDG] on the leniency bias in supervisors' ratings. RDG, as defined for this study, refers to a system that defines the percentage of subordinates that should fall into each rating category, which usually approximates a normal distribution (Ewenstein et al., 2016; Schleicher et al., 2009; Stewart et al., 2010).[4] However, unlike the more traditional forced distributions, RDG allows a supervisor to apply judgment to deviate from the provided guidance if warranted (Ewenstein et al., 2016; Schleicher et al., 2009; Stewart et al., 2010). As such, RDG is often perceived as fairer by subordinates and, thus, is becoming more popular than forced distributions in the practice (Mercer, 2013; Stewart et al., 2010). The downside to allowing more judgment is that RDG provides more opportunity for leniency to remain since there is less pressure and accountability to adhere strictly to a specific distribution (B. D. Blume et al., 2009; Olson & Davis, 2003; Stewart et al., 2010).[5]

Even without RDG, supervisors likely know that the injunctive norm (i.e., what one should do; e.g., Cialdini & Trost, 1998 Cialdini & Trost, 1998) is to prepare accurate subjective performance evaluations. Despite supervisors likely understanding they *should* prepare accurate ratings, there are many reasons why bias may still exist in their ratings, including a desire to manage impressions and present themselves positively to others. However, creating a formal RDG policy strengthens the injunctive norm for accuracy in the subjective performance evaluation process by explicitly communicating an expected distribution (Kaptein, 2015; McKinney et al., 2010; Shoemaker et al., 2020; Trevino & Nelson, 2014). I expect this strengthened injunctive norm for accuracy will decrease leniency bias, as compared to when RDG is absent. Indeed, prior research finds that when organizations provide supervisors with guidance on the rating distribution

[4] For example, an organization could provide RDG to supervisors stating that the expected distribution of subordinate rankings should be that 10% of subordinates *should* be rated below average, 80% *should* be rated moderate, and 10% *should* be rated above average (Bretz et al., 1992).

[5] As compared with a forced distribution where the supervisor *must* adhere to the guidance provided by the organization.

(both forced and RDG), there is an increase in differentiation and a decrease in leniency bias in subordinate ratings compared to when such guidance is absent (B. D. Blume et al., 2009; Grote, 2005; Scullen et al., 2005; e.g., Stewart et al., 2010). As such, my second hypothesis predicts that the use of RDG in the performance evaluation process will reduce leniency bias, as compared to when RDG is not provided. I expect that RDG will do so by making the injunctive norm of accurate ratings salient, thus reducing leniency bias.

My third prediction examines the use of PCCs and RDG together in an organization. In the presence of both a PCC and RDG, supervisors have two principal tactics for managing impressions. Tactic one would see supervisors continue to provide lenient ratings, despite the RDG, to present themselves as more capable managers, as predicted by my first hypothesis. Tactic two would see supervisors follow the communicated distribution to present themselves as willing to comply with the injunctive norms communicated by the RDG, in line with hypothesis two. These two tactics place opposing pressures on the leniency bias in a supervisor's ratings when that supervisor faces a PCC and RDG is provided. Thus, supervisors must decide how best to manage the impressions of their peers. To explore how these opposing forces jointly impact leniency bias, I consider how a supervisor may perceive what the social group (the PCC) approves of or disapproves of (the group's norms), as the most salient norm is the one most likely to impact the behaviour of individuals trying to manage impressions (Aronson et al., 2010; Bicchieri, 2006; Schneider, 1981; Shoemaker et al., 2020).

On the one hand, social comparison pressure created by a PCC will remain high, and thus, the desire to present oneself as a high-performing manager will remain strong regardless of the communication of RDG. On the other hand, the provision of RDG creates a precise reference point against which other members of the PCC can assess a supervisor's ability to rate subordinates accurately and adhere to group norms (Cialdini et al., 1991; Bicchieri, 2006). Furthermore, prior research has also found that when the desired outcome of an audience (i.e., the PCC) is known to individuals, they are more likely to conform to that desired outcome (Bonner, 2008; Lerner et al., 1999). Therefore, I anticipate a PCC to still induce upward pressure on ratings due to social comparison pressures and a desire to engage in impression management. However, I expect that the increased saliency of the group's injunctive norms, created by RDG, will weaken the effect of the PCC on leniency bias. Therefore, my third hypothesis is that the effect of a peer calibration committee on leniency bias will be weaker in the presence versus absence of RDG.

I test my hypotheses using a 2x2 between-subject experiment. I manipulate the presence versus absence of a PCC and the presence versus absence of RDG. Participants are assigned the supervisor role and instructed to rate a set of ten subordinates within their randomly assigned treatment condition. I compare, between conditions, the leniency bias that the supervisors display in the ratings they assign to the subordinate profiles. I first assess leniency bias across the entire population (*Average Leniency)* and do not find support for my hypotheses at the average group level. Next, I partition subordinates into groups based on their performance level - *Low, Moderate*, and *High Performers*. This partitioning is done on the basis of prior literature that suggests leniency bias may be asymmetrically applied, as supervisors may perceive a higher chance of conflict with lower-rated subordinates and thus may be more likely to provide low performers with higher ratings to avoid this conflict (B. D. Blume et al., 2009; Napier & Latham, 1986; Prendergast & Topel, 1993).

When analyzing by performance level, I find support for my first and second hypotheses within *Low Performers.* I find support that supervisors may display more leniency in ratings prepared in anticipation of a PCC, especially among low performers. As the increased bias appears to impact low-performers, this may create additional fairness concerns for moderate and high-performers, which could demotivate these subordinates. Additionally, it is essential to be aware of this potential increased leniency bias to ensure effective calibration within the PCC. I find support that RDG does have a main effect of reducing the leniency bias displayed among low performers. Further, using planned contrast testing, I find support for my predicted pattern of results for low performers. That is, the presence of a PCC has a main effect of increasing leniency bias, the presence of RDG has the main effect of reducing leniency bias, and when a PCC is present, the presence of RDG weakens the effect of PCCs on leniency bias. This finding indicates that RDG may be helpful in settings with a PCC. However, perhaps a stronger control, such as forced distributions, is required to mitigate more of the impact of PCCs on leniency bias.

Overall, my research offers several contributions to practice and theory. First, I contribute to the performance management literature by assessing the leniency bias displayed by supervisors anticipating participation in a PCC. My study is a meaningful extension of studies such as Bol et al. (2019), Grabner et al. (2020), and Demeré et al. (2019) that examine post-committee outcomes in settings where a calibration committee is always present. This study adds an examination of a PCC's impact on supervisors' initial ratings prepared in anticipation of a PCC as compared to those prepared without the expectation of a PCC. Understanding this pre-PCC impact is vital, as the quality of the ratings prepared for a PCC directly impacts the quality of the ratings post-PCC.

Second, I incorporate how RDG interacts with a PCC. In practice, RDG is a common control often used in conjunction with calibration committees (Bol et al., 2019; Demeré et al., 2019; e.g., Mercer, 2013). Given the predicted opposing effects on leniency bias, understanding how the two controls interact is important to inform their joint use in practice. I also provide an extension to prior studies that examine RDG in settings without a PCC (e.g., Stewart et al., 2010).

Third, my study instrument adds the ability to examine additional features. I demonstrate the ability to test social comparison in an online environment; this extends prior literature that examines social comparison in a laboratory environment (Hannan et al., 2013; e.g., Tafkov, 2013). I also offer insights into how impression management, social comparison, and norms are influenced by the presence of a PCC and RDG. Additionally, in contrast to most prior experimental literature that examines leniency bias by examing one or two subordinates, I create an instrument that allows for a comparison of leniency bias between low, moderate, and high performers. By studying a larger group of subordinates, I can better comment on the impact of biases on various levels of subordinate performance, and this greater breadth of performance levels expands the opportunities to test and understand the systematic nature of leniency bias.

The remainder of my thesis is organized as follows. In Chapter 2, I review psychology and accounting literature to examine subjective performance evaluations and the current understanding of leniency bias, PCCs, and RDG. In Chapter 3, I develop my predictions. In Chapter 4, I present my research design. In Chapter 5, I discuss the results of the hypotheses testing. Lastly, Chapter 6 presents my conclusions about this study's results.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In this chapter, I offer an overview of relevant psychology and accounting literature to provide an understanding of subjective performance evaluations, the impact of leniency bias on subjective evaluations, and two controls that may affect leniency bias. This chapter is organized into the following sections. Section 2.2 provides an overview of subjective performance evaluations, while section 2.3 discusses the biases present in the subjective performance evaluation process. Sections 2.4 and 2.5 discuss the impact of social comparison, impression management, and injunctive norms on the subjective performance evaluation process. Section 2.6 examines what is known about calibration committees' effects on subjective performance evaluation ratings. Section 2.7 discusses the impact of rating distribution guidance on subjective performance evaluations. Furthermore, section 2.8 discusses PCCs and RDG when used in conjunction. Finally, Section 2.9 concludes this chapter.

2.2 Subjective Performance Evaluations

In this section, I describe the use of performance evaluations and subjectivity in performance evaluations. I also discuss how subjectivity introduces bias into the performance evaluation process and the benefits and issues created by leniency bias, including why supervisors display leniency bias.

2.2.1 Subjective Performance Evaluations

Performance evaluations are a critical component of organizations' employee management systems. They are a tool used to recognize employees' achievements, provide feedback, and communicate the organization's goals (Bol, 2011; Moers, 2005). Evaluations also inform promotion, retention, and compensation decisions (Mercer, 2013; Moers, 2005). For many tasks, a supervisor can evaluate subordinates' performance using independently verifiable objective measures. However, for other tasks, objective measures are too costly or unavailable such that supervisors must subjectively evaluate subordinates' performance (Gibbs et al., 2004; Prendergast, 1999). Thus, subjectivity in performance evaluations allows for contracting on tasks when the subordinates' duties are not easy to contract explicitly or when output is not easy to measure objectively (Prendergast, 1999, p. 9). For example, subjectivity is useful when evaluating professional employees, such as auditors, consultants, managers, or information technology project staff, who do not have easily measured output quantity or quality (Bol, 2008; Gibbs et al., 2004).

Evaluations can contain subjectivity in three ways. First, evaluators can use subjective judgements to evaluate subordinate performance on tasks (Gibbs et al., 2004). For instance, creating objective measures for the quality of recommendations provided to clients or management is challenging. Therefore, a supervisor uses subjective judgment to evaluate subordinate performance on these types of responsibilities. Second, the subjective weighting of objective measures introduces subjectivity to evaluations (Gibbs et al., 2004). For example, a subordinate's assessment might include three objective measures: sales, customer satisfaction scores, and customer acquisition costs, but the weighting of these three measures may not be specified a priori for purposes of determining 'overall' performance. As such, a supervisor could subjectively choose to differentially weight the three measures when determining the final performance evaluation rating. Third, objective performance goals set at the beginning of a period could be subjectively adjusted ex-post (Gibbs et al., 2004). In this case, during the evaluation process, a supervisor could ex-post adjust (upwards or downwards) a subordinate's goal, such as a sales goal, to account for an uncontrollable factor that occurred in the period.[6] Including any of these three types of subjectivity in subordinates' performance evaluations can allow supervisors to provide more accurate, informative, and timely evaluations of a subordinate's contributions to the firm's value (Gibbs et al., 2004, p. 411). Additionally, subjectivity is often essential to evaluating the subordinate's contributions to an organization because subjectivity can provide crucial information that objective or quantitative output alone may not be able to measure (Bol, 2008; Gibbs et al., 2004; Grabner et al., 2020). I focus on the first of these three types of subjectivity: the use of subjective judgements to evaluate performance.

Subjectivity in evaluations can be crucial to evaluate subordinates' performance more appropriately. However, subjectivity may also reduce the accuracy of performance information, as supervisors may display bias when subjectively evaluating performance. Biased ratings are those ratings that systematically deviate from the "true" or "accurate" assessment (Bol, 2008, e.g., 2011; Ferris & Judge, 1991; Prendergast & Topel, 1993). The management research firm CEB finds that both nine out of ten supervisors and nine out of ten human resource leaders are dissatisfied with their organization's annual performance evaluations and believe the process does not yield accurate information (Wilkie, 2015).

2.3 Leniency Bias in Subjective Performance Evaluations

In this section, I will discuss how subjectivity may introduce leniency bias into the performance evaluation process and the benefits and issues created by leniency bias, including why supervisors display leniency bias.

[6] For example, the subordinate's goal could be adjusted downwards to account for the occurrence of an uncontrollable negative factor or (less frequently) increase the goal to ensure continued effort in the event of a 'windfall' event (uncontrollable positive event) that increases sales (Kelly et al., 2015; Gibbs et al., 2004).

2.3.1 Leniency Bias in Subjective Performance Evaluations

A specific bias that supervisors commonly exhibit when subjectively evaluating subordinate performance is leniency bias (e.g., Bol, 2011; Moers, 2005). Leniency bias is the tendency to give subordinates higher ratings than they deserve based on their ‘true’ performance (Saal & Landy, 1977). Moreover, research has shown that supervisors provide lenient ratings, even more so when ratings impact compensation or rewards (Jawahar & Williams, 1997; Milkovich & Newman, 1993).

Despite both supervisors and human resource professionals being aware that leniency bias can arise in subjective evaluations and understanding that it can lead to inaccurate and unbeneficial information, leniency bias persists (Wilkie, 2015). Leniency bias persists for four main reasons: 1) to avoid conflict with subordinates, 2) to reduce the effort involved with performance evaluations, 3) to use performance evaluations as an impression management tool, and 4) a lack of negative repercussions for displaying leniency bias.

2.3.1.1 Aversion to Conflict

Supervisors may be lenient because they want to avoid conflict and maintain positive relationships with their subordinates (Bol, 2011; Harris, 1994). Most subordinates tend to believe their performance is above average (Beer & Gery, 1972; Meyer, 1975) and overestimate their abilities compared to their supervisor’s assessment of their abilities (Harris & Schaubroeck, 1988; Shore & Thornton, 1986). This tendency can lead to conflict between supervisors and subordinates when the assessed rating is incongruent with a subordinate’s perception of their performance (Bernardin & Villanova, 1986; Friedrich, 1993; Lawler, 1990, etc. e.g., Murphy & Cleveland, 1991; Napier & Latham, 1986).

A conflict between supervisors and subordinates can occur at any level of subordinate performance; however, the risk for conflict is especially high when rating lower-performing subordinates (Napier & Latham, 1986). The tendency to overestimate one's abilities means that few, if any, subordinates will perceive their performance as below average. Therefore, a below-average rating is more likely to cause incongruency between a subordinate's perceived and assessed performance. Furthermore, the negative stigma and potential consequences associated with a below-average rating increase the likelihood that subordinates will react more intensely to a below-average rating. Thus, assigning a below-average rating—such as "needs improvement" or "not meeting expectations"—is more likely to harm the supervisor's relationship with that subordinate (Cohen-Charash & Spector, 2001; Colquitt et al., 2001; Greenberg, 1990; Napier & Latham, 1986) and potentially lead to uncomfortable confrontations (Barrett, 1966; Napier & Latham, 1986). Conversely, providing a low-performing subordinate with a lenient rating—such as "meets expectations"—will likely result in less conflict between supervisors and subordinates as it increases the congruence between the subordinate's perceived and assessed performance rating.

Though this effect may be stronger with lower-performing subordinates, the same principle holds for all performance levels. That is, providing less lenient ratings may increase conflict between the supervisor and the subordinates due to the incongruence between the supervisor's rating and the subordinate's perception of their own performance. Therefore, supervisors provide lenient ratings to subordinates out of a desire to avoid conflict with subordinates (Bernardin & Villanova, 1986; Friedrich, 1993; Lawler, 1990; Murphy & Cleveland, 1991; Napier & Latham, 1986).

2.3.1.2 Reduction of Effort

Supervisors may be lenient in an effort to limit the costly effort associated with preparing evaluations. For example, when anticipating assigning a rating lower than a subordinate may expect or want, supervisors may feel the need to spend more time on an information search to justify that lower rating to the subordinate (Arshad, 2020; Bol, 2008, 2011).[7] Therefore, some supervisors may be lenient to reduce the effort required to evaluate their subordinates.

2.3.1.3 Performance Ratings as Impression Management Tool

Supervisors may use lenient ratings as an impression management tactic to present themselves more favourably to superiors and peers. Impression management is "an attempt by one person to affect the perceptions of her or him by another person target" (Schneider, 1981, p. 25). The desire to manage impressions causes people to seek to maintain, improve, or alter how others perceive them (Goffman, 1959; Leary & Kowalski, 1990; Webb et al., 2010). Supervisors may seek to manage impressions by providing subordinates with higher ratings to signal better leadership skills and a well-functioning department (Bol et al., 2019; Rosaz & Villeval, 2012).[8]

2.3.1.4 A Lack of Negative Repercussions for Displaying Leniency Bias

Supervisors rarely face negative repercussions for inaccurate (biased) evaluations. Subjective performance evaluations require the application of judgment, which inherently creates ambiguity about rating accuracy. This ambiguity makes it challenging to evaluate the accuracy of a supervisor's ratings and hold supervisors accountable for leniency.

[7] Relatedly, the subordinate's incongruent perception of their performance in relation to their rating increases the occurrence of requests for justification from subordinates (Barrett, 1966; Bol, 2008; Napier & Latham, 1986).

[8] Impression management is an notable aspect of behaviour within a performance evaluation system, as such, section 2.4 explores impression management in more depth.

Additionally, supervisors are not usually the residual claimant of the organization’s profits (Grund & Przemeck, 2012; Prendergast, 1999); that is, supervisors do not typically bear the cost associated with any compensation increase, bonus, or other monetary rewards resulting from lenient ratings (Prendergast, 1999). Thus, supervisors, not bearing any potential financial costs, might continue to provide lenient ratings to subordinates, despite an organization’s desire for accuracy.

2.3.2 The Benefits and Costs of Leniency Bias

Understanding leniency bias requires examining both its positive and negative effects on subordinate and organizational performance. Prior research has shown some positive effects of leniency bias in ratings. For example, providing higher ratings reduces the risk of confrontation between the supervisor and the subordinate. This risk of conflict is especially true for lower-performing subordinates who would otherwise receive low ratings based on their ‘true’ performance (Napier & Latham, 1986). Additionally, there is evidence that lenient ratings may cause subordinates to appreciate their supervisor more (Spence & Keeping, 2011) and that leniency may increase subordinate motivation (Bol, 2011).

On the other hand, much prior research shows that leniency bias is detrimental to subordinate performance and organizational decision-making. Lenient ratings provide subordinates with inaccurate information about their performance, thus weakening the link between effort and incentives, which can lead to less effort and suboptimal subordinate performance (Baker et al., 1988; Bénabou & Tirole, 2005; Bol, 2008; Fang & Moscarini, 2002). Leniency bias also weakens the communication of the organization’s desired goals, which may lead to poor effort direction (Baker et al., 1988; Bol, 2008, 2011). For example, providing a low performer with a rating indicating they have ‘met expectations’ suggests that they do not need to change their behaviour even though such subordinates may need to perform tasks differently or put in more effort. Therefore, leniency bias can decrease subordinate performance through inadequate communication of expectations and goals (Baker et al., 1988; Bol, 2008).

Additionally, when leniency bias is present in evaluations, subordinates may negatively perceive the fairness of their ratings relative to their perceptions of their peers' performance (Bol, 2011).[9] For example, a high-performing subordinate might be resentful if one of her low-performing subordinate peers receives a higher rating than she believes is warranted. Consequently, this may lead to a negative view of subsequent decisions based on these ratings, such as compensation and reward decisions (Bol, 2008; Bol et al., 2019; Ewenstein et al., 2016). On the whole, a negative perception of fairness may decrease subordinate motivation (Colquitt & Chertkoff, 2002; Erdogan, 2002).

Leniency bias also may cause problems for decision-making in organizations as it may result in the compression of subordinate ratings (Milkovich & Newman, 1993; Moers, 2005). This compression may be caused by the asymmetric application of leniency bias, with lower performers benefiting more from the leniency bias as supervisors perceive more chance of conflict with lower performers. Thus, supervisors may be more likely to provide low performers with higher ratings to avoid this conflict (as previously discussed; Napier & Latham, 1986; Prendergast & Topel, 1993). Rating compression causes issues in organizations as decision-makers may not be able to clearly distinguish between subordinates to make the best compensation, promotion, and retention decisions for their organizations (Bol, 2008; Jawahar & Williams, 1997; Mercer, 2013; Prendergast, 1999).

[9] This is predicated on subordinates being aware of other subordinate's ratings. This would be the case in a company that makes public (internally) relative performance information, but in organizations that do not publicly share this information the availability of the knowledge would rely on more informal peer-to-peer information sharing. Alternatively, a subordinate could rely on their perception of the relative ratings based on communicated expectations or a lack of perceived consequences for poor performance.

Overall, leniency bias and the resulting compression of ratings can cause subordinate effort intensity and direction issues and may reduce the ratings' informativeness for decision-makers in the organization. In addition, theory and evidence point to several reasons supervisors tend to display leniency in their ratings in the absence of controls that intend to reduce these tendencies.

2.4 Social Comparison and Impression Management in Performance Evaluations

This section discusses accounting and psychology literature regarding the performance evaluation process, social comparison, and the supervisors' need to manage the impression of peers and higher-level managers regarding their leadership abilities.

2.4.1 Social Comparison and Impression Management

Social comparison theory offers that people seek to evaluate their own abilities by comparing themselves to others (D. Brown et al., 2007; Festinger, 1954; Suls & Wheeler, 2000; Tafkov, 2013). People use these social comparisons to form their self-image, aiming to avoid negative feelings and maintain a positive self-image (Beach & Tesser, 1995; Lazarus, 1991; R. H. Smith, 2000; Tesser, 1988). To achieve a better self-image, individuals may compete to outperform others (D. Brown et al., 2007; Garcia & Tor, 2007).

Prior research finds three main factors help strengthen the feeling of social comparison and, consequently, the behaviours that individuals exhibit in response. First, the sense of social comparison is stronger when the tasks performed are similar among group members and when the group members are of similar ability (Festinger, 1954; Garcia & Tor, 2007; Harkins & Jackson, 1985; Tafkov, 2013). Second, the feeling of social comparison is stronger when differences in performance on the compared task can be attributable to controllable factors, such as ability and effort, and not to uncontrollable factors, such as luck (Festinger, 1954; Garcia & Tor, 2007;

Goethals & Darley, 1977; Tafkov, 2013). Third, the feeling of social comparison is stronger when others in the comparison group are important enough to the individual to care about their opinions, thus important enough to evoke social comparison (Suls & Wheeler, 2000; Tafkov, 2013). Workplaces, especially teams within these workplaces, feature these three factors resulting in feelings of social comparison.

Moreover, a performance evaluation process requires a supervisor to assess subordinates' performance, typically on facets related to effort, ability, or both. Performance management systems also formally define the subordinate group a supervisor evaluates, and a supervisor's peer group is often partially defined by their subordinate group. For example, a supervisor's peer group may consist of supervisors who manage teams in the same department, teams with similar job levels, teams who perform similar tasks, or another such subset of subordinates (Bol et al., 2019; Mercer, 2013). Therefore, a performance management system may enhance social comparison, particularly if it emphasizes the comparison of subordinates between supervisors.

Individuals may seek to actively manage impressions to present a positive image and deal with social comparison pressures in the workplace. The target of impression management can be any or all of the superiors, peers, or subordinates (Webb et al., 2010; Yukl & Falbe, 1990). However, the ability to engage in impression management also requires an opportunity to do so. As discussed, organizational performance management systems can enhance a supervisor's feeling of social comparison. Further, subordinate ratings can create a specific point of comparison among the supervisors' peer group. As such, supervisors can use their subordinates' performance ratings as a medium for social comparison to achieve a positive self-image (Rosaz & Villeval, 2012). Therefore, supervisors may have the desire and the opportunity to use these ratings to demonstrate superior managerial skills and engage in impression management.

To develop a positive self-image, supervisors will seek to appear as highly capable managers to their peers (D. Brown et al., 2007; Garcia & Tor, 2007; Tafkov, 2013). Prior research finds support that supervisors provide lenient ratings to their subordinates to signal their managerial ability to others, including their peers and higher-level managers (Bol et al., 2019; Ilgen et al., 1981; Merchant & Van der Stede, 2017; Rosaz & Villeval, 2012). Providing lenient ratings to signal managerial ability is an example of using the performance ratings of subordinates to manage impressions. Specifically, higher subordinate ratings may signal better managerial skills and a well-functioning department (Bol et al., 2019, p. 13; Rosaz & Villeval, 2012). Longnecker et al.'s (1987) findings from interviews with 60 executives indicate that supervisors inflated their subordinate ratings when they were going to be reviewed by others outside the department. These findings are consistent with using lenient ratings as an impression management tactic.

Overall, supervisors play a dual role in the performance evaluation process. Supervisors not only evaluate their subordinates, but their superiors also evaluate them against their peers (i.e., other supervisors of a similar level). As supervisory skills are a component of a supervisor's role, demonstrating a well-performing department through high subordinate ratings could be a way to manage impressions with their superiors and peers (Bol et al., 2019; Ilgen et al., 1981; Merchant & Van der Stede, 2017; Rosaz & Villeval, 2012). Consequently, those supervisors who wish to achieve high social standing among their peers and high ratings themselves are likely to engage in impression management in the performance review process. Therefore, given that social comparison and impression management exist in—and are potentially increased by—performance management systems, it is vital to understand the impact of a chosen performance evaluation system design.

2.5 Injunctive Norms in the Subjective Evaluation Process

Injunctive norms "specify what people approve and disapprove within the culture and motivate action by promising social sanctions for normative or counter normative conduct" (Reno et al., 1993, p. 104). More simply, an injunctive norm is a perception of how one *should* behave (Cialdini & Trost, 1998). For example, supervisors likely know the injunctive norm within a performance evaluation context is to prepare *accurate* subjective performance evaluations. Injunctive norms can develop naturally or through specific communication by an organization, such as a code of conduct (Kaptein, 2015; McKinney et al., 2010; Shoemaker et al., 2020; Trevino & Nelson, 2014). For example, Tayler and Bloomfield (2011) find that formal controls influence people's sense of injunctive norms.[10,11] Cialdini et al. (1991) also find that norms can affect behaviour, even when not associated with any direct economic consequences.

Furthermore, Tayler and Bloomfield (2011) note that people may simultaneously hold multiple injunctive norms and that contextual features, such as controls, often determine the norm activated in a situation. Norm activation theory comes from work by Endler (1993) and Cialdini & Trost (1998), which finds that norms are activated depending on which norm is the most salient at any given time. To summarize prior research, it finds that people can hold multiple norms at any one time, and the norm most likely to govern behaviour is the one activated by the current situation, such as the context created by the environment or task (Cialdini & Trost, 1998; Endler, 1993; Tayler & Bloomfield, 2011).

[10] Tayler and Bloomfield refer to this type of norm as a 'personal norm'. A personal norm holds the same underlying basis as an injunctive norm, i.e., that it represents one's own sense of what *should* be the appropriate behaviour (Cialdini et al., 1990).

[11] There are two main types of norms – injunctive and descriptive norms. Injunctive norms are beliefs about what one *should* do, and descriptive norms are beliefs about what people *actually* do (Cialdini et al., 1991; Cialdini et al., 1990). Both personal and injunctive norms are used in prior literature to describe internal beliefs about what one *should* do. Additionally, the discussion in Tayler and Bloomfield (2011) refers to personal norms as a type of injunctive norm, specifically juxtaposing personal norms against descriptive norms.

In the subjective performance evaluation process, injunctive norms may develop in many ways. As one such example, human resource departments might provide instructions that include a performance rating scale and information about how the organization's management will use the rating outcomes (i.e., promotion, retention, and compensation decisions; Mercer, 2013; Wilkie, 2015). [12] These instructions then either create or strengthen injunctive norms within the performance evaluation process, such as the understanding that the supervisor *should* prepare accurate ratings (Aronson et al., 2010; Pelfrey & Peacock, 1991; Shoemaker et al., 2020). Despite this, research consistently shows that other environmental and task factors may override an injunctive norm for accuracy, resulting in supervisors still exhibiting leniency bias (Bol, 2008, e.g., 2011; Ferris & Judge, 1991; Moers, 2005; Prendergast & Topel, 1993). As discussed above, these overriding factors include aversion to conflict, avoiding effort, a lack of repercussions for biased ratings, and the desire to manage impressions. (Arshad et al., 2020; Bol, 2008, 2011; Grund & Przemeck, 2012; Harris, 1994; Hecht et al., 2020; Prendergast, 1999).

In summary, while supervisors likely understand that they *should* prepare accurate ratings, there are many reasons why bias may still exist in their ratings, including a desire to manage impressions and present themselves positively to others. Therefore, using controls in the performance evaluation process may help increase supervisors' rating accuracy and reduce leniency bias by making the injunctive norm of accurate ratings salient.

2.6 Calibration Committees

An organization's dissatisfaction with the bias in performance evaluations leads many to find alternatives or to establish various controls to increase rating accuracy. For example,

[12] Performance rating scales can be numeric (e.g., 3, 4, 5), alphabetic (e.g., a, b, c), or narrative (e.g., below expectations, meets expectations, above expectations). Five-level performance rating scales are common, but other variants are also used (Bretz et al., 1992; SHRM).

organizations may choose to supplement or replace annual performance reviews with more frequent feedback or feedback from more parties (i.e., 360-degree feedback); alternatively, they may add controls such as additional training on performance evaluations for supervisors, improved rating scales, or implementing calibration committees (Demeré et al., 2019; Landy & Farr, 1980; D. E. Smith, 1986; Wilkie, 2015; Woehr & Huffcutt, 1994). This paper will first explore one such control; the impact of calibration committees on performance evaluation bias. This section defines calibration committees and discusses the purpose of using such committees and how they are employed. In addition, I summarize the current research regarding the impact of calibration committees on leniency bias.

2.6.1 Calibration Committees

A calibration committee is a group of supervisors whose task is to collectively discuss and possibly adjust subordinate performance ratings (Bol et al., 2019). Organizations widely use calibration committees, and many consultants promote them as a best practice to help reduce leniency bias (Albert, 2017; Hastings, 2011; Mercer, 2013; Risher, 2014). The 2013 Mercer Global Performance Management Survey Report (GPMSR) finds that 56% of organizations employ calibration committees, and of those organizations, 93% use the committee to review and discuss performance ratings (Mercer, 2013).[13,14]

[13] A broad range of organizations engage in activities to calibrate their ratings. The GPMSR finds that companies in the consumer goods, banking, and durable manufacturing industries are slightly more likely to impose controls that mandate specific calibration processes or use a forced distribution as opposed to guidelines for rating distributions (Mercer, 2013, p. 9). The GPMSR (Mercer, 2013) includes responses from over 14 different industry categories including manufacturing and consumer goods, technology, and financial and professional services. The survey also "includes responses from performance management leaders of 1,056 organizations representing 53 countries around the globe. The organizations surveyed varied in size from fewer than 1,000 employees to more than 10,000 employees and represent a wide variety of industries and structures (for-profit, non-profit, government)" (Mercer, 2013, p. 1). The 2013 GPMSR includes 40% North American Companies, 21% European Companies, and 27% Asian Companies.

[14] Other issues discussed by calibration committees include compensation (31%), succession planning (29%), and learning and development (26%) (Mercer, 2013).

Organizations typically use one of three types of calibration committees: (1) peer-level calibration committees [PCCs], comprised of peer supervisors (Arshad et al., 2020; Bol et al., 2019); (2) higher-level calibration committees [HCCs], comprised of supervisor's superiors (Demeré et al., 2019); or (3) a combination of peer-level and higher-level supervisors (Grabner et al., 2020). While data on each committee type's prevalence is unavailable, anecdotal evidence suggests PCCs are more common in practice than other forms of committee composition (Bol et al., 2019). Moreover, academic studies (Arshad et al., 2020; e.g., Bol et al., 2019; Lillis et al., 2017) and practitioner literature (Caruso, 2013; Hastings, 2011; Risher, 2011; e.g., Sammer, 2008) mostly describe calibration committees as peer-level rather than HCCs or calibration committees with combined levels.

Prior research presents some insights regarding the impact of using a calibration committee on the extent of leniency bias. For example, Bol et al.'s (2019) field study finds that, on average, the performance ratings of subordinates are lower (i.e., less leniency bias) *after* discussion by the PCC compared to *before* discussion by the PCC (Bol et al., 2019). In addition, Arshad et al. (2020) find that supervisors gather additional information when preparing for a PCC discussion. Previous research has shown that in some circumstances, increased supervisor effort when assigning ratings may reduce bias (Bol, 2008, 2011; Bonner, 2008; Moers, 2005). Therefore, it is possible to assume that all else being equal, an increased information search signals increased effort, which could result in less bias. However, Arshad et al. (2020) do not specifically examine if less bias occurs due to the additional information gathering. Instead, they focus on other aspects of the PCC process, including information sharing and rating adjustments occurring *within* the PCC meeting. Thus, it is still unknown how a PCC, even with the additional information search shown by Arshad et al. (2020), affects a supervisor's pre-committee leniency. Thus, more examination of this pre-committee behaviour and leniency is necessary.

In summary, there is some evidence that PCCs reduce the bias in ratings from pre-to-post-committee reviews (e.g., Bol et al., 2019). However, how supervisors' anticipation of participating in PCC will impact the extent of leniency bias in ratings prepared for the committee's consideration during an evaluation process is not yet known. As the quality of the inputs limits the quality of any output, understanding the impact of a PCC on the ratings prepared by supervisors in anticipation of a committee's review is essential to understanding a calibration committee's impact on leniency bias.

2.7 Rating Distribution Guidance (RDG)

This section reviews RDG and its impact on leniency bias in subjective performance evaluations. I discuss the types of rating distribution systems, the known effects of these systems on leniency bias in performance evaluation ratings, and the benefits and disadvantages of rating distribution systems.

2.7.1 Rating Distribution Guidance

Providing supervisors with guidance on the rating distribution for subordinates can, in place of or in addition to other controls, increase the saliency of the injunctive norm regarding accurate performance ratings. The broad definition for any form of rating distribution is—a control within a performance evaluation system that assists with rating and ranking subordinates (Stewart et al., 2010, p. 168). More specifically, RDG, as used in this study, refers to a system that defines the percentage of subordinates that should fall into each rating category, which usually approximates a normal distribution (Ewenstein et al., 2016; Schleicher et al., 2009; Stewart et al., 2010).

Guidance on rating distributions takes two primary forms: *forced* distributions and *expected* distributions.[15] As the term implies, a *forced* distribution is one where the rating outcomes *must* adhere to a particular set of parameters (McBriarty, 1988). For example, the required (forced) distribution could be that 10% of subordinates must be rated below average, 80% must be rated moderate, and 10% must be rated above average. Forced distributions reduce leniency bias by holding the supervisor accountable to a prescribed rating distribution (B. Blume et al., 2013; Grote, 2005; e.g., Schleicher et al., 2009; Scullen et al., 2005). Alternatively, *expected* distributions (RDG hereafter) might identify a similar distribution but use an approach that allows a supervisor to apply judgment to deviate from the guidance if they believe doing so is warranted (Stewart et al., 2010).

Prior research finds that when organizations provide supervisors with guidance on the rating distribution (both forced and RDG), there is an increase in differentiation (i.e., a decrease in centrality bias) of subordinate ratings compared to when such guidance is absent. (B. D. Blume et al., 2009; Grote, 2005; Scullen et al., 2005; e.g., Stewart et al., 2010). For example, Blume et al. (2009) find much greater differentiation among subordinates with a forced distribution system. However, they also find that a forced distribution system is not without risks. For example, individuals had less attraction to work at the organization based on the conditions associated with the forced distribution system (e.g., consequences of poor performance, reward differentiation, etc.).

Additionally, Blume et al. (2009) find that since a forced distribution system increases differentiation and reduces leniency bias, those most impacted are the low performers. This asymmetric impact is because these low-performing individuals are more likely than moderate and

[15] Prior research uses various terms to describe RDG, including: expected distributions, forced distribution, forced ranking systems, bell curves, group ordering, or normal distributions (Bol et al., 2019; Demeré et al., 2019; Stewart et al., 2010). Each term has slightly different usage, but all focus on the idea of a prescribed distribution for the subordinate ratings.

high performers to receive a lower rating under a forced rating distribution system than the rating received when no rating distribution system is in place (B. D. Blume et al., 2009). For example, Olson and Davis (2003) use Ford Motor Company as an example of success with rating differentiation when using a forced distribution system. Before implementing a forced distribution system at Ford, 98% of managers received a 'fully meeting expectations' rating, whereas post-implementation guidance required increased differentiation among the managers, reducing some of the managers' ratings (Olson & Davis, 2003). These findings indicate that RDG can be effective in reducing leniency bias.

Despite reducing leniency and increasing differentiation, forced distribution systems have fallen out of favour in organizations. One key reason is that forced distribution systems can lead subordinates to perceive ratings as unfair (B. D. Blume et al., 2009; Olson & Davis, 2003; Stewart et al., 2010). This negative perception arises since a *forced* distribution does not allow supervisors to exercise discretion in assigning more subordinates to a particular (e.g., higher) rating category, even if they believe it is warranted (Schleicher et al., 2009). Therefore, some subordinates may feel their rating is artificially low due to the forced nature of the distribution (Schleicher et al., 2009). This negative fairness perception may demotivate the subordinates receiving lower ratings, especially if compensation and performance ratings are linked (Olson & Davis, 2003; Schleicher et al., 2009; Schrage, 2000). Consequently, this negative effect on motivation may negatively impact subordinate performance and, thereby, organizational performance.

RDG can provide similar benefits as a forced distribution system by directing supervisors' efforts to meet the communicated distribution. The trade-off with the RDG approach is that the ratings do not need to strictly adhere to the guidance, so leniency can still occur. Nevertheless, RDG allows supervisors to apply judgment, which often means subordinates perceive the distribution as fairer than a forced distribution (Stewart et al., 2010).

The opportunity for judgement with the RDG approach can allow supervisors to focus on the subordinate's task performance directly, rather than just trying to fit the subordinate into a specific category on the suggested distribution (e.g., "exceeds expectations") compared to their peers. This room for judgment potentially enables each subordinate to be assigned fairer ratings. However, as RDG requires a not-so-strict adherence to the suggested distribution, it still provides the opportunity for leniency as there is less pressure and accountability to adhere to a specific distribution.

On the whole, RDG is more prevalent in practice than forced distributions (Mercer, 2013), and thus, my study will focus on RDG. According to Mercer (2013), of the companies surveyed globally, 55% use RDG, whereas 30% use forced distribution, with some industries using each at even higher rates. In addition, *forced* distributions have faced legal challenges in cases where a company was not careful with its implementation and policies (Bates, 2003; Osborne & McCann, 2004; Stewart et al., 2010). Finally, field studies examining calibration committees provide evidence that organizations use RDG in conjunction with calibration committees (Bol et al., 2019; Demeré et al., 2019; Grabner et al., 2020). However, the research designs of these studies do not allow for examination of the impact of RDG on the calibration committee's behaviour as there is no condition where RDG is not present (Bol et al., 2019; Demeré et al., 2019; Grabner et al., 2020). Therefore, studying the combined effects of RDG and PCCs is needed to inform both research and practice.

2.8 Rating Distribution Guidance in Environments where Peer Calibration Committees are Present

As previously discussed, organizations often use RDG in conjunction with calibration committees, and prior field studies have examined the outcome of calibration committees in settings where organizations use RDG (Bol et al., 2019; Demeré et al., 2019; e.g., Grabner et al., 2020). Therefore, examining the interaction of these two performance evaluation controls is vital to fully understanding how PCCs function in organizations.

In their field study, Bol et al. (2019) examine an organization with a PCC that also provides RDG. When all subordinates discussed in the PCC are considered as one group (regardless of the subordinate's supervisor), Bol et al. (2019) find that for the entire group of subordinates discussed in a PCC meeting, the post-PCC rating distribution conforms to the provided RDG. However, Bol et al. (2019) also note that communicating RDG to a calibration committee may not necessarily induce conformation to the expected distribution within a supervisor's specific group of subordinates post-PCC.

Several possible reasons may explain why a specific supervisor's ratings post-PCC may not conform to RDG. For example, Bol et al. (2019) find a supervisor's political power in the organization may reduce another supervisor's willingness to disagree with their ratings in the PCC.[16] Bol et al. (2019) also explore other factors, such as how some supervisors might have a greater willingness to 'fight harder' for higher ratings at PCC meetings, or conversely, how some supervisors' aversion to conflict with other PCC members may reduce a supervisor's desire to fight for their subordinates' ratings. However, Bol et al. (2019) primarily focus on the changes

[16] Moreover, the findings also show that a supervisor's alliances or network among PCC members may allow them to achieve the ratings they desire (Bol et al., 2020).

between pre-and-post-PCC ratings and how the interplay between committee members may influence rating adjustments.

Despite this focus, Bol et al. (2019) also note that communication of RDG to PCC members may not necessarily induce adherence to the expected distribution in the ratings a supervisor prepares in anticipation of a PCC. However, Bol et al. (2019) do not directly examine the impact of a PCC on ratings prepared in anticipation of participating in such a meeting. Moreover, Bol et al. (2019) find that some individual supervisors' *post*-PCC ratings show non-conformity to the RDG. Accordingly, it may be inferred that if a supervisor's post-PCC ratings do not conform to the RDG, then it is likely that the ratings prepared for the PCC discussion would also not conform to the RDG. Therefore, further exploring how PCCs and RDG interact to affect leniency bias in pre-PCC ratings is warranted.

2.9 Conclusion

This chapter reviewed the relevant psychology and accounting research examining leniency bias in subjective performance evaluations and the role of calibration committees and RDG as ways of potentially addressing such bias. Overall, the existing literature suggests supervisors often exhibit leniency bias when preparing subjective performance evaluations. However, despite supervisors likely understanding that providing accurate ratings is desirable from the organization's perspective, strong unintended incentives often exist to provide lenient ratings. In addition, while consultants continue to promote calibration committees as a best practice for reducing leniency bias, the limited research on their use and consequences leaves open the question of the impact of PCCs on the ratings a supervisor prepares in anticipation of the committee's review and possible rating adjustments. Additionally, while prior research provides some understanding of how RDG impacts leniency bias when used in isolation, the combined effects of

RDG and PCCs have not been studied. I will examine these issues in detail in Chapter 3, where I develop my hypotheses.

CHAPTER 3: DEVELOPMENT OF HYPOTHESES

3.1 Introduction

In this section, I use theories of social comparison, impression management, and social norms to develop hypotheses about supervisors' behaviour when conducting performance evaluations. Specifically, I investigate how the anticipation of participating in a PCC affects leniency bias in supervisors' ratings. In addition, I examine how RDG affects the bias displayed in those pre-PCC ratings.

I organize this chapter as follows. Section 3.2 develops my prediction regarding the effect of PCCs on leniency bias. Section 3.3 develops my prediction regarding the impact of RDG on leniency bias. Section 3.4 develops my prediction about how RDG moderates the effect of PCC on leniency bias. Finally, section 3.5 summarizes the chapter.

3.2 The Effects of a Peer Calibration Committee on Leniency Bias

Prior research on the use of calibration committees has focused either on the adjusted ratings arising from the calibration committee review and discussions or on pre-PCC deliberation information search behaviour (Arshad et al., 2020; Bol et al., 2019). However, I expect another characteristic of PCCs will likely impact the leniency bias in a supervisor's *pre*-PCC ratings, social comparison. As previously discussed, a PCC is a group of peer supervisors who meet to discuss their respective teams' performance evaluation ratings (Bol et al., 2019; Grabner et al., 2020; Mercer, 2019). As such, a PCC creates a forum, or audience, which may increase social comparison of the managerial ability among committee members. Since individuals use social comparison to achieve a positive self-image and appear superior to others (Beach & Tesser, 1995;

Festinger, 1954; i.e., Tafkov, 2013), I expect this increased social comparison shifts more of a supervisor's focus to impression management in front of their peers.

Subordinate ratings are a mechanism a supervisor can use to manage impressions since having higher ratings may create the appearance of having a well-managed, high-performing team (Grabner et al., 2020; Ilgen et al., 1981; Merchant & Van der Stede, 2017; Rosaz & Villeval, 2012). One of the roles of a supervisor in an organization is to provide guidance and support to subordinates; therefore, it is plausible that subordinate performance can, to some degree, be associated with a supervisor's managerial abilities (Grabner et al., 2020). For example, a field study by Dineen et al. (2006) finds that supervisors who provide guidance to subordinates can induce higher levels of organizational citizenship and less errant behaviour, resulting in better subordinate performance. Since the quality of the guidance provided by supervisors is unlikely to be directly observed by others in an organization, the rating a subordinate receives may signal to others the quality of that subordinate and, indirectly, the supervisor's ability to guide and support that subordinate. A supervisor may have impression management concerns that can lead to leniency bias (Bol et al., 2019; Grabner et al., 2020; Longenecker et al., 1987; Rosaz & Villeval, 2012). Indeed, there is evidence that supervisors will provide lenient ratings to their subordinates evidence that supervisors provide lenient subordinate ratings to present themselves as better managers to their superiors if there is a payoff attached to the ratings (i.e., their own evaluation, compensation increases, bonuses, etc.).

Research has shown that individuals are more likely to engage in impression management when they are in the presence of others or acting within a group (Leary & Kowalski, 1990; Schlenker, 1980). Individuals may also feel pressure to conform to group norms and expectations, which can further increase the likelihood of impression management behaviours (Leary &

Kowalski, 1990). When individuals are concerned about how they are perceived by others, they may engage in impression management strategies to enhance their self-presentation (Jones & Pittman, 1982). This may include providing exaggerated information about their performance or abilities. For example, in a study conducted by Dittmar et al. (2014), participants were asked to imagine they were taking part in a performance evaluation at work. They were then asked to rate the extent to which they would engage in impression management behaviours, such as exaggerating their accomplishments. The results showed that participants who reported higher social comparison concerns were more likely to engage in impression management behaviours, especially when they were part of a group (Dittmar et al., 2014). The tendency to provide exaggerated information about one's performance or abilities is known as self-enhancement. Self-enhancement refers to the tendency of individuals to present themselves in a more positive light than is warranted by reality (S. E. Taylor & Brown, 1988). Research has shown that individuals who are more concerned with impression management tend to engage in more self-enhancement (Paulhus, 1991). Moreover, individuals may be more likely to engage in self-enhancement when they perceive a threat to their self-esteem, such as when they experience social comparison concerns.

In addition, group dynamics can create a competitive environment where individuals are motivated to outperform their peers in order to gain recognition and respect (Klein & Kunda, 1992). A study by Brown and Levinson (1987) found that individuals were more likely to engage in self-promotion when they believed that their reputation was at stake, such as when they were being evaluated by others. This suggests that the desire to engage in impression management may be particularly strong in situations where an individual's performance is being evaluated by others.

As discussed in Chapter 2, prior research identifies three factors that lead to increased social comparison effects within a group: (1) the task performed and the ability of the group members must be similar, (2) differences in task outcomes can be attributed to differences in ability and effort, and (3) the comparison group must be important enough to create social comparison concerns (Festinger, 1954; Garcia & Tor, 2007; Harkins & Jackson, 1985; Tafkov, 2013). All three of these are present and salient within a PCC. First, supervisors who comprise a PCC are typically similar in rank (i.e., peers), such that each supervisor has the similar task of guiding, motivating, and monitoring subordinates. Second, a supervisor's ability to guide, motivate, and monitor subordinates can affect subordinates' performance (Dineen et al., 2006). Third, PCCs are comprised of supervisors who typically know each other and work together in an organization (Bol et al., 2019; Demeré et al., 2019; Grabner et al., 2020). Therefore, the PCCs' members are likely important enough to evoke greater social comparison.[17]

However, this argument is not without tension. First, a supervisor could potentially use the PCC process to shift responsibility for a lower rating onto a PCC and, thus, mitigate some of the conflicts that may arise when they provide a subordinate with a low rating. If this occurs, it may lead to less leniency bias in ratings.

Second, instead of providing lenient ratings to self-promote, a supervisor could try to demonstrate their ability by providing the most accurate ratings possible. However, it is not clear that providing more accurate ratings would self-promote the supervisor within a PCC setting. In the absence of a benchmark, or other such information, it is difficult for other PCC members to directly judge the accuracy of a supervisor's ratings. As such, a supervisor is more likely to favour

[17] Furthermore, the supervisors on a committee may also have a common superior responsible for evaluating all the supervisors on the committee, thereby increasing the comparison pressure even more. However, I do not operationalize supervisors having a common superior in my study. By not including this factor, I can examine the impact of PCC participation without the confound of a superior's expectations.

the more salient option of providing lenient ratings that present their team's best attributes and, thus, their skills as a supervisor.

In summary, theory and prior literature support that a supervisor anticipating reporting to a PCC will have greater impression management concerns than a supervisor with final authority. Specifically, a PCC creates an audience that makes social comparison among peers highly salient. This heightened saliency is further increased by the nature of a PCC, whose function is to share and discuss each supervisor's subordinates' performance ratings (Bol et al., 2019; Mercer, 2019). As such, a PCC inherently creates an environment where each supervisor's subordinates are compared. This increased social comparison heightens impression management desires. Consequently, relative to those with final authority, supervisors reporting to a PCC will have stronger incentives to self-promote their managerial ability by providing lenient ratings for their subordinates (Figure 1).

H1: Leniency bias will be higher when supervisors anticipate reporting subjective performance ratings to a peer calibration committee than when they have final authority over those ratings.

3.3 The Effect of Rating Distribution Guidance

As discussed in Chapter 2, RDG specifies the percentage of subordinates that should fall into each rating category. For example, RDG can expressly state that an organization's expectations are: 10% of subordinates perform at a below-average level, 80% perform at an average level, and 10% perform at an above-average level (e.g., Bretz et al., 1992). However, unlike forced distributions, RDG allows supervisors to apply discretion to deviate from the guidance when judged as appropriate (Stewart et al., 2010). As such, the mere presence of RDG does not guarantee complete adherence to the distribution. Instead, by explicitly communicating an expected rating distribution, RDG strengthens the injunctive norm for accuracy within a subjective performance evaluation process compared to when RDG is absent (Kaptein, 2015; McKinney et al., 2010; Shoemaker et al., 2020; Trevino & Nelson, 2014). This strengthened injunctive norm for accuracy helps to reduce the leniency bias in the subjective evaluation process.

Though, as discussed in Section 2.3.2, an injunctive norm for accurate ratings likely exists in most organizations (Aronson et al., 2010; Pelfrey & Peacock, 1991; Shoemaker et al., 2020), the increased formalization provided by RDG strengthens the saliency of the injunctive norm for supervisors (Trevino & Nelson, 2014). This strengthened norm will, in turn, increase the obligation supervisors feel to provide more accurate ratings (Cialdini & Trost, 1998; Trevino & Nelson, 2014), thereby reducing leniency bias. In addition, the introduction of RDG also formalizes and communicates specific information on what the organization deems to be an 'accurate' rating range. This formalization provides a benchmark against which others in the organization could assess whether a supervisor's ratings align with the organization's rating norms, further cementing the saliency of this injunctive norm for accuracy.

In summary, RDG specifies the expected rating distribution (outcome) desired by the organization, increasing the injunctive norm's saliency for accuracy and leading to a reduction in leniency bias.

H2: Leniency bias will be lower when organizations give supervisors rating distribution guidance than when organizations do not give supervisors rating distribution guidance.

3.4 The Interactive Effect of Peer Calibration Committees and Rating Distribution Guidance

As discussed in the development of H1, a PCC's presence increases social comparison and a supervisor's desire to engage in impression management, which is expected to *increase* leniency bias. On the other hand, H2 predicts that RDG will strengthen the saliency of the injunctive norm for accurate ratings, which is expected to *reduce* leniency bias. For my third hypothesis, I consider how these two controls jointly impact leniency bias. To that end, I first consider how the perception of what the social group (i.e., the PCC) approves or disapproves of (i.e., the PCC's injunctive norms) is likely to impact the behaviour of supervisors trying to manage impressions (Schneider, 1981).

When both a PCC and RDG are present, supervisors have two dominant tactics for managing impressions. Tactic one would see supervisors continue to provide lenient ratings, disregarding the RDG, to present themselves as more capable managers (as argued in the development of H1). Tactic two would see supervisors follow the RDG to present themselves as willing to comply with the group's injunctive norm (as argued in the development of H2). Thus, when facing a PCC and having been provided RDG, a supervisor must choose how they will manage their peers' impressions, with the two available tactics placing opposing pressures on the level of the leniency of their ratings.

To predict which effect will most substantially impact supervisor behaviour, I first consider prior research on norm activation. Previous research shows that the most salient cues govern a specific norm's activation (Bicchieri, 2006; Schwartz, 1977; Tayler & Bloomfield, 2011). The more salient a norm, the stronger its effect on one's behaviour (Cialdini et al., 1991; de Araújo, 2014; e.g., Reno et al., 1993). For example, Cialdini et al. (1991) take the commonly understood injunctive norm against littering and activate it through environmental factors. Some participants are in an environment that contains garbage on the ground, and some are in a clean environment. Cialdini et al. (1991) find that a tidier environment induces individuals to litter less often. The activation effect is even more substantial when the participant observes an experiment confederate exhibiting non-littering behaviour (Cialdini et al., 1991). While they assess that all participants understand the societal injunctive norm against littering and the descriptive norm that many people still litter, the specific activation of this societal injunctive norm induces more non-littering behaviour.

Further, prior research also finds that when norms provide differing tactics for managing impressions, the expectations of how others will behave typically offer the most salient information about the appropriate course of action (Aronson et al., 2010; Bicchieri, 2006; Shoemaker et al., 2020). For example, Shoemaker et al. (2020) examine individuals' use of company computers for personal tasks during work hours. The code of conduct outlines that individuals should not engage in personal tasks on their computers at work (i.e., social media, online shopping, internet browsing, etc.). By doing so, the organization provides formal communication about how an individual *should* behave and strengthens the saliency injunctive norm not to use the internet for personal reasons at work (Shoemaker et al., 2020). However, individuals who observe their peers engaging in personal internet use at work are more likely to engage in personal internet use themselves since they can justify that even though they *should not* do so, everyone else does. So, individuals can also justify using the computer for personal tasks (Shoemaker et al., 2020). Following similar reasoning, in my setting, I expect that the behaviours supervisors engage in to manage impressions will depend on what is more salient; the desire to self-promote by providing lenient ratings or the injunctive norm to provide accurate ratings.

To determine which tactic will have the most saliency when both RDG and PCCs are present, I first consider the nature of the PCC. As discussed in Section 3.2, a PCC creates an environment that enhances social comparison and creates strong desires to manage impressions. Therefore, I expect the desire to present oneself as a high-performing manager will remain strong even with the communication of RDG. Next, I consider that when an organization uses RDG and PCC in combination, the communication of RDG provides an explicit reference against which other members of the PCC can assess a supervisor's ability to rate subordinates accurately. As discussed in Section 3.3, this direct communication strengthens the saliency of the injunctive norms regarding rating behaviours for the supervisors participating in a PCC (Shoemaker et al., 2020; e.g., Trevino & Nelson, 2014). Therefore, RDG's communication of a precise reference against which to assess a supervisor's ratings may decrease a supervisor's willingness to be lenient. This would occur as closer adherence to RDG would demonstrate to the group (the PCC) a supervisor's willingness to adhere to the group's norms (Bicchieri, 2006; Cialdini et al., 1991).

Furthermore, prior research has also found that when the desired outcome of an audience (i.e., the PCC) is known to individuals, they are more likely to conform to that desired outcome (Bonner, 2008; Lerner & Tetlock, 1999). Following the known desired outcome allows individuals to reduce their cognitive effort and demonstrate their willingness to comply with the desired outcome (Bonner, 2008; Lerner & Tetlock, 1999). RDG provides readily available information about acceptable rating decisions (Mero et al., 2007; e.g., Tetlock, 1992). As such, supervisors may more closely adhere to the RDG and provide less lenient ratings to reduce their cognitive effort by adhering to the communicated injunctive norms strengthened by the RDG.

Considering both sides, I expect a PCC to induce upwards pressure on ratings (i.e., leniency bias). This upward pressure on ratings is caused by: (1) the increased social comparison pressure created by the direct audience of peers and (2) the resulting impression management desires to present oneself as a high-performing supervisor. However, I expect that this effect will be weaker when RDG is present. RDG weakens the effect of the PCC on leniency bias by providing readily available information regarding the injunction norms for acceptable ratings (i.e., the desired outcome) of the audience (i.e., the PCC).

In summary, when an organization uses a PCC and RDG together, theory predicts that the RDG will weaken the impact of a PCC on leniency bias. (Figure 3 and

Figure 4), leading to my final hypothesis:

H3: The effect of a peer calibration committee on leniency bias will be weaker in the presence versus absence of rating distribution guidance.

CHAPTER 4: RESEARCH METHOD

4.1 Introduction

I use a 2 x 2 between-subjects experiment to test my hypotheses. I manipulate the presence versus absence of a PCC and the presence versus absence of an RDG. Participants are assigned the role of supervisor and instructed to provide ratings to a set of ten subordinates within a randomly-assigned treatment condition. My primary dependent variable is the leniency bias supervisors display in their subordinate ratings.

The remainder of the chapter is organized as follows. Section 4.2 provides an overview of the experimental task and main task components. The participants in the experiment are reviewed in Section 4.3. Next, section 4.4 explains the main experimental task, including the general instructions and the main scenario details presented to participants. Section 4.5 discusses the subordinate profiles participants use to assess subordinate ratings. Section 4.6 discusses the key dependent variables and process measures collected in the experiment. Lastly, this chapter concludes in Section 4.7.

4.2 Experimental Design Overview

In my experiment, participants take on the role of a supervisor and rate a set of ten subordinates.[18] Appendix A provides an overview of the experiment's flow, as further described next:

1) Participants begin by receiving identical general instructions (Appendix B). These instructions summarize the key details of the task, inform participants they will be taking

[18] I discuss additional details on the participants selected for this experiment in Section 4.3.

on the role of a supervisor, and inform participants they will be rating a team of their subordinates within the context of annual performance evaluations.

2) Participants are presented with one of four scenarios outlining the key features of the experimental condition to which they have been assigned (Appendix C).[19]

3) Participants complete four knowledge check questions tailored to their assigned condition (see knowledge check questions in Appendix C). These knowledge check questions highlight the key aspects of the assigned condition to help ensure that participants have absorbed the proper information from the presented scenario. Participants only proceed once they correctly answer all knowledge check questions.

4) Participants are presented with ten subordinate profiles and participants are asked to provide a subjective performance rating for each subordinate (see example profile in Appendix D). Subordinate profiles are presented on different screens, one after the other. [20]

5) Participants are asked a series of post-experiment questions after the main task to examine process measures and collect relevant demographic information (Appendix E). Specifically, I seek to capture the social comparison and impression management pressures participants experienced, as well as participants' beliefs regarding the injunctive norms in the performance evaluation process.[21]

[19] I further discuss the key elements of these scenarios in Section 4.4.

[20] Section 4.5 further discusses the subordinate profiles.

[21] See further discussion of process measures in Section 4.6.2.

4.3 Participants

I use online labour pool participants recruited from Prolific in my experiment.[22] Using online participants allows me to effectively match the participants to my experiment's task (R. Libby et al., 2002). My task does not necessarily require specific expertise (e.g., high technical knowledge), as is required in some studies (e.g., audit judgment studies). However, to understand a subordinate rating task, including pressures supervisors may face in a performance evaluation process, it is important that participants have supervisory experience. For example, someone with prior supervisory experience will have a better understanding of discussing a performance evaluation with a subordinate and a better understanding of the injunctive norms present when preparing the evaluation. Therefore, I feel supervisory experience provides a foundation for participants to relate to the scenario in my experiment and to provide greater generalizability of my tested theories (e.g., Bailey et al., 2011; T. Libby et al., 2004) and using a prolific sample allows me to select participants with this prior experience.

Prolific collects and stores specific demographic data regarding participants for screening participants for studies (*Prolific*, n.d.). Examples of screening items include supervisory experience, stock market experience, and education level (*Prolific*, n.d.). As Prolific collects screening criteria independently from any specific research study (*Prolific*, n.d.), participants do not know the most beneficial answer to a question about their background (e.g., education). Therefore, participants are less likely to provide a false answer to gain access to a specific study (Palan & Schitter, 2018). I use screening criteria on Prolific to pre-screen participants for supervisory experience. At the end of the study, I also obtain secondary confirmation of participants' supervisory experience with the question, "*How many employees have you supervised*

[22] For my pilot studies I recruit participants from MTurk. Both MTurk and Prolific participants have successfully replicated findings from prior research (Peer et al., 2017).

at one time?”.[23] As such, I anticipate that participants will have the knowledge and skills required to perform a subordinate performance evaluation task.

Numerous accounting studies have used online labour pools, such as MTurk and Prolific, as a participant source (i.e., Callan et al., 2017; Marreiros et al., 2017).[24] Moreover, online workers are increasingly being used for management accounting experiments as they are: (1) representative of the general population (A. M. Farrell et al., 2017; Garrow et al., 2020) and (2) good proxies for non-expert workers (A. M. Farrell et al., 2017). Online labour pool participants, including Prolific, have successfully replicated findings from prior research (Peer et al., 2017). I chose Prolific for my experiment as the workers on this platform produce less unusable responses compared to both other online labour markets and university subject pools (Peer et al., 2017). Further, I conduct a Pilot Test (Appendix F) to assess the suitability of online participants for my specific task, including inducing social comparison in an online setting and finding the pool to be suitable.

As my hypotheses do not rely on variation in pay rates or other compensation differences, all participants receive a fixed compensation of £2.00 for participation in my study.[25]

In summary, I recruit participants from the online labour platform Prolific. I pre-screen for supervisory experience, education of at least an undergraduate degree, and participants living in the United Kingdom, Canada, the United States, Australia, or New Zealand.

[23] Prolific users do not know they were initially screened on supervisory experience, as the study is only presented to those who have previously answered affirmatively that they have supervised subordinates. While supervisory experience is mentioned in the study's recruiting information, it is likely that participants either did not focus on this information or forgot about this requirement by the time they answered demographic questions at the end of the study.

[24] In addition, prolific has been used successfully in other fields such as economics (e.g., Marreiros et al., 2017) and psychology (e.g., Callan et al., 2017).

[25] Prolific requires payment of at least £5.00/hour and recommends £7.00/hour. Therefore, remuneration of £2.00 for a task taking approximately 15 minutes would be the equivalent of £8.00/hour, making this a fair wage for this participant group. Exchange rates from the approximate time the study was conducted (March 2021) equates £2.00 with $3.52 CAD based on the Bank of Canada Exchange Rate from the first week of March 2021. An approximate completion time of 15 minutes results in an hourly pay rate of $14.08 CAD.

4.4 General Instructions and Scenarios

4.4.1 General Instructions

All participants first review general instructions intended to introduce them to the experiment (Appendix B).[26] These general instructions highlight that participants are assuming the role of a team supervisor, that they will be determining subordinate performance ratings for the year, and other basic experimental information. These instructions are identical for all experimental conditions.

4.4.2 Base Scenario

In all conditions, participants next receive the main scenario (Appendix C). All conditions have the same scenario base, but key sections are tailored for each experimental condition to which participants are randomly assigned. Specifically, all scenarios inform participants they are engaging in an annual performance review process and will be assessing ten of their subordinates. Participants are also introduced to the ten-point scale they will use in the rating portion of the task (from Poor Performance – 1 to Exceptional Performance – 10). Other essential information given to ***all*** participants includes:

(1) *"The firm believes all supervisors should provide accurate ratings as doing so provides important feedback to employees about their performance."*;
(2) *"Assume after you complete the evaluations, you will be responsible for discussing them with your employees."*
(3) *"You would then submit their evaluations to HR for inclusion in each employee's file.";*
(4) *"A summary of your team's ratings will also be sent to your supervisor as part of his/her resources for assessing your own performance."*; and
(5) *"The firm believes how an employee performs reflects both the employee's ability and effort and the supervisor's ability to bring out the best in their employees."*

[26] The general instructions are adapted from (Bailey et al., 2011). Additionally, to use language more familiar to participants, I use the word "employee", throughout the experiment rather than subordinate.

This information helps to establish a common baseline for all participants regarding the performance evaluation process to help reduce information differences, and, thereby, potential noise that could be caused by participants' real-world experiences.[27] These details also help add realism to the presented scenarios. Using more realism in scenarios can help increase observed effects, increase external validity, increase the generalizability of results, and help create a balance between natural settings and the representations made by scenarios (Aguinis & Bradley, 2014; Hughes & Huby, 2002; Taylor, 2006; Wason et al., 2002).

I include the details provided in the above statements for specific reasons. I explicitly state the organization desires accurate ratings and why, point (1), to help create a shared and salient understanding of the organization's accuracy norms (Bicchieri, 2006; Schwartz, 1977). Point (5) discusses that subordinate ratings reflect supervisors' ability, to help reinforce the relationship between supervisor and subordinate performance. It is important to include both these points since an experimental setting is devoid of the usual pressures and norms developed in real-world organizations through the daily interactions with subordinates, peers, superiors, and general organizational tasks and policies. As this information is provided in all conditions, the effects of this information should not significantly affect inferences from hypotheses testing; further, any effects would tend to bias against (rather than towards) finding support for my hypotheses.

Points (2), (3), and (4) highlight that performance reviews will be discussed with subordinates, filed by HR, and reviewed by superiors. As discussed in Section 2.3, prior research finds these stakeholders may affect supervisors' leniency bias. Therefore, I include these details to add realism, establish the evaluation process flow, and provide context about these key stakeholders in the performance evaluation process. Additionally, in the conditions where the

[27] Section 4.3 discusses participants including why I elect to use participants with past supervisory experience.

supervisor has final authority (i.e., where a PCC is absent), the mention of filing with human resources helps to establish that there is no review of the ratings before filing.[28]

For the base condition, where neither a PCC nor RDG is present (control; NoPCC and NoRDG), these are the only key instructional details received before viewing and rating the subordinate profiles.

4.4.3 PCC Present - Scenario

My first independent variable is the presence or absence of a PCC. When a PCC is absent, the scenario provides no information about a PCC process. When a PCC is present, I incorporate details into the scenario to establish the PCC process, the PCC's composition, and the PCC's goals. Specifically, when a PCC is present, I tell participants to:

> *"[…]assume after you complete the employee ratings, the next step will be to meet with a group of 4 other supervisors in your department. This committee will review the ratings of each employee and calibrate ratings across supervisors."* (Appendix C).

And

> "*The goal of this committee is to reduce any differences in employee ratings across the supervisors*" (Appendix C).

These two sentences establish the overall purpose and size of the PCC and reinforce the primary goal of reducing rating differences (i.e., calibrating subordinate ratings).

Participants also receive more specific details in the scenario about the other members of the PCC (Appendix C), which are intended to enhance the relationship, and, thereby, social comparison between participants. This enhanced relationship will allow for a more robust test of theory when using participants from an online labour market (i.e., Prolific). To my knowledge, at

[28] When a PCC is present, information on the PCC is given prior to this statement, and so is established as part of the process with the filing with human resources something that occurs after the main evaluation process.

the time of conducting this experiment, there were no previous online experiments within this context that also relied on social comparison and impression management theory. However, given the outbreak of the Covid-19 pandemic in March 2020, it was not possible to conduct an in-person experiment. As such, I pilot test the key details provided about the PCC, to assess the strength of social comparison they can induce in online participants (Appendix F). My pilot test supports that social comparison can be induced in online participants.

These PCC member details were designed based on the three factors for a strong relationship between social comparison and behaviour; 1) task similarity, 2) differences being due to factors the individual can control, and 3) the comparison group is important (Festinger, 1954; Garcia & Tor, 2007; Harkins & Jackson, 1985; Tafkov, 2013). Furthermore, these details are based on features of PCCs found in practice (Bol et al., 2019; Demeré et al., 2019; Mercer, 2013). My discussion of Pilot Test One in Appendix F provides a detailed breakdown of these details and the intended social comparison factors they address. Based on the results from Pilot Test One, I include the following details to strengthen social comparison in my online scenario (Appendix C):

> *"Assume the following about the calibration committee members:*
> - *They all manage teams in your department*
> - *They all have similar work experience to yours*
> - *They all have employee teams that are similar to yours*
> - *They all present their employee ratings to the committee for review*
> - *You frequently interact and work with each of them*
> - *You care a great deal that they think you are a good supervisor"*.

4.4.4 RDG Present - Scenario

The second manipulated variable is the presence versus absence of RDG. To determine the appropriate RDG provide, I follow the assumption from prior research that the 'true' distribution of subordinates' performance ratings across an organization will follow a normal distribution

(Bretz et al., 1992; e.g., Ewenstein et al., 2016; Moers, 2005).[29] Under an assumption of normal distribution, when performance rating distributions deviate from normality the cause is attributed to bias (Aguinis, 2009; Bol, 2008; Bol & Smith, 2011; Moers, 2005; O'Boyle Jr & Aguinis, 2012, p. 82; Schneier, 1977). For this study, I adopt the view that a normal distribution represents the "true" or "accurate" assessment of a group of subordinates, as is common in accounting research (Bol, 2011; Bol & Smith, 2011; e.g., Grabner et al., 2020; Moers, 2005), as well as, in organizational behaviour and human resource management research (e.g., Bretz et al., 1992; Holzbach, 1978; Morris et al., 2015; von Sydow et al., 2019).[30] As such, the RDG provided in my study and my assessment of leniency bias will follow a normal distribution.

When RDG is absent, the scenario provides no information about the expected subordinate rating distribution. When RDG is present, the scenario informs participants:

> *"Human Resources has provided guidance approved by the CEO that on average across all supervisors and departments, 20% of subordinates should be rated 8-10; 20% of subordinates should be rated 1-3; with the remaining 60% rated using the range of 4 to 7"* (Appendix C).

This guidance establishes participants' expectations of the distribution and emphasizes, by evoking the CEO's approval of the RDG, that adhering to this distribution is important to the firm.

[29] Organizational behaviour and human resource management research have long held the assumption that subordinate performance follows a Gaussian (normal) distribution. This assumption underlies most statistical analyses and theory development in these fields. (Hull, 1928; O'Boyle Jr & Aguinis, 2012; Schmidt & Hunter, 1983; Tiffin, 1947). Some researchers and practitioners believe that actual subordinate performance does not follow a normal distribution (e.g., Bernardin & Beatty, 1984; Saal et al., 1980), with some researchers suggesting alternative distributions, such as a Paretian (power-law) distribution (O'Boyle Jr & Aguinis, 2012; e.g., West et al., 1995). Nonetheless, a Gaussian distribution remains the standard assumption for performance evaluation distribution in performance evaluation literature and practice (O'Boyle Jr & Aguinis, 2012). Gaussian distributions follow a standard normal distribution, which assumes that the mean and standard deviation are stable. In contrast, Paretian distributions assume that means and standard deviations are not stable. Instead, these distributions follow power law, where one variable varies as a relative proportion (power) of the other variable. These distributions characterized by unstable means, infinite variance, and a greater quantity of extreme events (O'Boyle Jr & Aguinis, 2012).

[30] For argument's sake, if instead a Paretian distribution is adopted, the issue of leniency bias (further discussed in section 2.2.3) would only be more pronounced in a Paretian distribution than in a Gaussian (normal) distribution. Gaussian distributions follow a standard normal distribution, which assumes that the mean and standard deviation are stable. In contrast, Paretian distributions assume that means and standard deviations are not stable. Instead, Paretian distributions follow power law, where one variable varies as a relative proportion (power) of the other variable. These distributions characterised by unstable means, infinite variance, and a greater quantity of extreme events (O'Boyle Jr & Aguinis, 2012). Thus, in this case, a Gaussian (normal) distribution not only falls in line with the majority of practice and research but is also a more conservative test of the presence of leniency bias.

I use the terms ‘guidance’ and ‘on average’ to allow participants to use judgment in determining the actual rating distribution. This is consistent with guidance that may be provided in organizations when an RDG and not a forced distribution is used. After the presented scenario, I ask all participants a series of knowledge check questions that correspond to the scenario they view to ensure understanding before continuing to the main task of rating subordinates’ profiles.

4.5 Subordinate Profiles

After the scenario, each participant views a series of subordinate profiles [*Profile(s)*] (Appendix D). All *Profiles* include six evaluation categories, and participants are asked to provide a final overall rating for each (section 4.5.1 discusses the *Profile* presentation). I present ten *Profiles*, one after the other, on separate screens (section 4.5.2 discusses the number of *Profiles*). Three *Profiles* present information consistent with lower-performing subordinates [*Low Performers*], four *Profiles* present information consistent with moderately-performing subordinates [*Moderate Performers*], and three *Profiles* present information consistent with higher-performing subordinates [*High Performers*] (section 4.5.3 discusses this *Profile* distribution). The *Profile* presentation order deliberately distributes *Low, Moderate, and High Performers,* with specific focus to distribute the *Low* and *High Performers* between the first and second half of the profile set to avoid either a front or back-loading of these performance levels. The *Profiles* are presented in the same order for every participant.[31]

4.5.1 Subordinate Profile Presentation

As mentioned above, each *Profile* includes six evaluation categories, and participants are asked to provide a final overall rating. Specifically, *Profiles* show six performance criteria for each

[31] I do not predict an order effect of the presentation of profiles. Were an order effect to occur, the consistent order presentation of the Profiles should mean that it will affect all conditions, and therefore, would not bias hypothesis testing.

of the ten subordinate profiles: 1) Cooperative Behaviour, 2) Leadership, 3) Business Development and Networking Skills, 4) Organizational Skills, 5) Initiative, and 6) Time Management (See example of a final *Profile* in Appendix D). I inform participants in all conditions that the six performance criteria categories should be evenly weighted when evaluating a subordinate's performance. This information helps ensure a common understanding of the informational weight of the six categories to avoid noise due to differing interpretations of how to use the information provided. I conduct a second pilot test to examine the exact design of the presentation of *Profiles;* the details of this testing can be found in Appendix G. I pilot test the *Profiles'* presentation to gain assurance that the *Profile* presentation is understandable to participants and does not create excess noise while still allowing for variation in ratings provided. To test this, I present participants with three different *Profile* presentations to examine if participants would: 1) provide differentiated ratings between subordinate profiles, 2) that the ratings assigned ***within*** each *Profile* fall within a reasonable distribution based on the intended subordinate profile performance level. Testing indicates that all three profile presentations are suitable. Selection of the *Profile* presentation is, therefore, based on which presentation best meets the aforementioned criteria and has the most straightforward presentation of information. Therefore, *Profile* presentation two is selected (Appendix G – Panel C).

As part of my consideration in designing my profiles, I consider that prior literature has shown that respondents may have differences in response styles to scaled experimental and survey questions. These differences in response style have been shown to impact the number that participants select on a rating scale and, therefore, cause biased results (e.g., Wetzel et al., 2016; Bolt & Johnson, 2009; Baumgartner & Steenkamp, 2001). Most studies examining this response bias suggest statistical methods for eliminating the effects of bias in response. However, the goal of this study is to capture leniency bias. As such, I need to resolve this issue through other means to help ensure any bias shown is caused only by the factors manipulated in my experiment and not the general response bias present in participant scale responses. Therefore, to help reduce unintended noise and unreasonable variation in ratings within *Profile* I added an additional design feature. Each *Profile* presents six categories of rating information to allow for subjectivity in the final rating; however, I provide participants with some preliminary information on a reasonable range for each of the six categories. Instructions inform participants to assume that the 'blue bars' on the rating scale represent their own preliminary rating ranges for these categories. Each blue bar covers a range of three ratings on a 10-point scale (Appendix D). This range of three rating scores creates a more standardized reference point but also provides an opportunity to capture variation between participants to test my hypotheses[32]

To further reduce response bias caused by an individual's scale interpretation, I also use a second commonly suggested method. This method is to provide labels along the rating scale, and not just at the endpoints, to add more context to the numbers from the scale (e.g., Wetzel et al., 2016; Bolt & Johnson, 2009; Baumgartner & Steenkamp, 2001). These two features should help reduce extraneous noise in response to the rating scales.

4.5.2 The Number of Subordinate Profiles

In addition to the presentation of each individual *Profile*, I also consider the number of subordinate *Profiles* to include in my instrument. To make this determination, I begin by considering the nature of leniency bias. Leniency bias is the systematic tendency to provide more lenient ratings than warranted by a subordinate's performance (Saal & Landy, 1977). Thus, I seek

[32] For example, to minimize the possibility that two participants would assess an 'average' Profile differently due to differences in their beliefs on what rating represents an 'average' performer —such as a rating of 6 for one participant versus a rating of 8 for another participant.

to include enough *Profiles* to capture a more systematic view of leniency, which contrasts prior research that typically examines only one or two subordinates (e.g., Bol & Smith, 2011; Arshad et al., 2020). By including more *Profiles* I also allow for the opportunity to examine how a PCC and RDG may differentially affect subordinates at different performance levels (i.e., *Low*, *Moderate* and *High Perf*ormers). Next, I consider that participants need to rate a sufficient number of *Profiles* to have an opportunity to apply RDG meaningfully, including deviating from RDG if desired. For example, in practice, supervisors do not just place a single subordinate on a rating distribution without the context of also needing to place other subordinates on that distribution. By using more *Profiles*, I can better reflect this reality to participants. Furthermore, I consider that using more *Profiles* may also decrease the potential that a demand effect is created by providing RDG. For example, if I provide only three *Profiles*, a participant provided RDG may feel an obligation to assign one subordinate to each performance level.[33] However, as the number of *Profiles* increases, participants may feel they can apply more judgment in determining each *Profile's* rating, even if their ratings result in deviation from the RDG. Considering the aforementioned factors, I include ten *Profiles* in my instrument. Without being too onerous, ten *Profiles* will allow for the examination of both the systematic tendency toward leniency bias and the potential differential application of leniency bias at different subordinate performance levels.

4.5.3 The Performance Level Distribution of Subordinate Profiles

After setting the number of *Profiles* at ten, I next consider, 1) the alignment between the RDG and the distribution of *Profiles* **between** each performance level (i.e., *Low, Moderate, and*

[33] By presenting a stark comparison of three RDG categories and three subordinate profiles participants could infer that the desired response is to place one participant in each category (Orne, 1962; Sears, 1986; Zizzo, 2010; Iyengar et al., 2011). This demand effect has been raised as a particular concern in an online labour pool (such as MTurk) due to participants' potential experience with other studies and their desire to have their responses accepted on the platform by researchers to maintain their MTurk quality rating (Berinsky et al., 2012; Goodman et al., 2013; Krupnikov & Levine, 2014).

High Performers) and 2) the more precise distribution of *Profiles* **within** each performance level. I seek to find a balance between these two objectives across the set of *Profiles*. On the one hand, the distribution of *Profiles* needs to be such that participants **could** rate *Profiles* precisely in line with the provided RDG. On the other hand, the distribution of *Profiles* needs to be such that participants **could** provide lenient ratings. Specifically, a participant should not feel they need to artificially drop a participant too far below or above the performance level presented by the *Profile*. Instead, when considering both the distribution **between** and **within** each performance level, the *Profiles* should offer participants enough flexibility to justify either 1) ratings that align with the RDG, or 2) ratings that are more lenient than the RDG.[34]

Striking this balance presents both a more conservative test of my hypotheses and increases the likelihood that a rating distribution with a negative skew (i.e., a distribution skew that sees more values concentrated on the right—higher—side of the distribution) is the result of participants providing lenient ratings. To illustrate this consider a counterexample; suppose participants evaluate a set of *Profiles* that ‘accurately’ fall in a distribution that has a negative skew (i.e., a lenient distribution). In such a circumstance, one would reasonably expect to see a negative skew to the distribution of assigned ratings. However, in this case, a negative skew might not demonstrate leniency bias. Instead, this negative skew might demonstrate increased rating accuracy and willingness to apply judgment to deviate from RDG. Such a circumstance would make any inferences regarding leniency bias and my hypothesized conditions challenging to validate. In contrast, by creating a set of *Profiles* that could reasonably follow the RDG provided, I bias against findings of a higher negative skew. This creates a test that is both more conservative

[34] If participants do not feel the presented *Profiles* could reasonably adhere to the RDG they may either: 1) ignore the RDG completely as it is too unrealistic, or 2) feel that the study wants them to adhere to the RDG despite it being unrealistic to do so, thus creating a potential demand effect. Alternatively, if *Profiles* do not provide enough flexbility participants may not have any oppourtunity to display leniency bias.

and creates a setting where I can more easily draw out conclusions related to my hypotheses. Thus, I create greater internal validity and comfort that negatively skewed ratings are indicative of leniency bias.

Given the above discussion, I design the set of *Profiles* to present all participants with three *Low Performer* subordinates, four *Moderate Performer* subordinates, and three *High Performer* subordinates. However, within each of the *Low Performer* and *High Performer* groups, I incorporate subordinates that are close to being *Moderate Performers*. This design follows practice where some supervisors might need to 'inaccurately' assign a subordinate's rating if they want to comply precisely with the range from the guidance. It also allows for a reasonable defence of both lenient ratings and ratings that follow RDG.

To further illustrate this, consider the following. To comply precisely with the RDG in the study instrument, a participant would need to rate two subordinates rated as low-performing, six as moderate-performing, and two as high-performing (Appendix C). However, the designed *Profiles* provide three low-performing, four moderate-performing, and three high-performing *Profiles*, with profiles present on the edge of low-to-moderate and moderate-to-high. Therefore, a participant could reasonably rate two subordinates as low performers, six as moderate performers, and two as high performers: precisely matching the RDG. However, a participant could also find a reasonable basis to be more lenient with some of the *Profiles*. Thereby, a participant with a tendency to display leniency bias would not need to unreasonably deviate from the information presented in the *Profiles* to also rate two subordinates as low performers, five as moderate performers and three as high performers (or one as low, six as moderate and three as high, or other such lenient distributions). Such ratings create a negative (lenient) skew to the participant's rating distribution. However, considering each *Profile* individually, a participant could focus more on the more positive information presented (confirmation bias) and, thus, expect they can reasonably justify a more lenient rating.

To summarize, I design a *Profile* set comprising ten subordinates following a normal distribution. This allows for an examination of the systematic nature of leniency bias and provides the opportunity to assess leniency bias at different performance levels. Additionally, to allow for a choice between the precise application of RDG or the demonstration of leniency bias, I design a *Profile* set with a distribution of three low-performing subordinates, four moderate-performing subordinates, and three high-performing subordinates, with one of the low-performing and one of the high-performing subordinates closer to the moderate-performing subordinates than the others.

4.6 Dependent Variables

My primary dependent variable is leniency bias. I calculate my main measure of leniency bias as the rating assigned by participants less the preliminary rating midpoint average from each of the six categories (see Appendix H).[35] Once a leniency bias measure is calculated for each *Profile*, I also create a variable that captures the average of the ten *Leniency Bias* measures *[Average Leniency]*. Further, I partition the ten *Profiles* into three categories: below-average (*Low Performers*), average (*Moderate Performers*), and above-average performers *(High Performers)*.

4.6.1 Creating a Leniency Bias Measure

Before conducting my analysis to test my hypothesis, I calculate a measure of leniency bias [*Leniency Bias*]. To accomplish this, I first calculate a non-biased midpoint rating [*Subordinate Midpoint Rating*] for each *Profile*. To find the *Subordinate Midpoint Rating* for each *Profile*, I first

[35] For example, the preliminary rating midpoints for *Subordinate 1*, as illustrated by Appendix H, are 6, 6, 8, 7, 8, and 5. Therefore, the average of these preliminary rating midpoints is 6.67. Next, suppose a participant provided *Employee 1* with a rating of 8. Then the leniency bias for *Subordinate 1* for this participant is calculated as: 8-6.50 = 1.5. See further discussion in Section 0.

calculate the midpoints of each provided preliminary range for all six categories on a *Profile* (Appendix H). For example, 'Employee 1' presents a preliminary rating range of five to seven for the "Cooperative Behaviour" category (Appendix D); therefore, the midpoint for the "Cooperative Behaviour" category 'Employee 1' is six. Next, I take these midpoints for each of the six categories on the *Profile* and calculate the median. This median becomes the *Subordinate Midpoint Rating* for that *Profile*. I repeat these steps for each of the ten *Profiles* in the study instrument. See Table 1, Panel D for the *Subordinate Midpoint Rating* for each *Profile*.

I then use the *Subordinate Midpoint Rating* to partition the *Profiles* into the performance level groupings discussed in Section 4.5.3: three low performers [*Low Performers*], four moderate performers [*Moderate Performers*], and three high performers [*High Performers*]. Appendix H provides summary statistics of the *Subordinate Midpoint Ratings* and the sorting of the *Profiles* into performance levels. This partitioning allows me to assess leniency bias within different levels of subordinate performance.

Separately examining each performance level gives a more thorough understanding of the impact of PCCs and RDG.[36] *High Performers,* for example, are more likely to experience a ceiling effect as their ratings can only be assessed so high. However, there is more room to demonstrate leniency bias with *Low Performers*; thus, I would expect the effects of leniency bias to be more significant in the *Low Performers'* group (Bol, 2011). In addition, a supervisor seeking to provide lenient ratings to manage peer impressions is unlikely to want subordinates at a below-average performance rating, since that could signal poor management skills (e.g., poor selection, training, or monitoring skills). In contrast, *Moderate Performers* and *High Performers* already fall into

[36] Low, moderate, and high performers are grouped based on the two subordinates with the lowest 'midpoint mean rating' (low), the six subordinates with 'midpoint mean rating' in the middle of the group (moderate), and the two subordinates with the highest 'midpoint mean rating' (high). This distribution is built into the subordinate profiles based on the RDG provided to subordinates.

average and above-average rating levels; which do not carry as much negative stigma. As a result, *Moderate Performers* and *High Performers* are less likely to induce as much leniency bias. Thus, separating subordinate profiles into *Low Performers, Moderate Performers,* and *High Performers* for analysis allows for a more thorough examination of leniency bias.[37]

4.6.2 Process Measures

I collect process measures in a post-experimental questionnaire to help examine my main theoretical mechanisms. I use seven-point Likert scales to collect process measures capturing a participant's social comparison concerns, impression management concerns, injunctive norm beliefs, and supplemental items (Appendix E).

First, I endeavour to capture a participant's social comparison and impression management concerns post-experiment. To capture each participant's social comparison concerns, I adapt three social comparison questions from Tafkov (2013) and Hannan et al. (2013) (Appendix E – Panel A – Questions 1-3).[38] I use two questions (two questions as per Webb et al., 2010) for my theoretical Impression Management construct (Appendix E – Panel A – Questions 4 and 5).

Second, I seek to capture a participant's injunctive norms, as they relate to performance evaluations. To evaluate this, I ask participants two questions to understand the injunctive norms they hold post-experiment regarding the distribution of ratings and the accuracy of ratings (Appendix E – Panel A – Questions 16-17). These two questions most directly relate to the goals

[37] Most prior experimental studies examine only one or two subordinate profiles (e.g., Bol & Smith, 2011; Arshad et al., 2020). As such, grouping by subordinate level is not a factor in these research designs. However, this prior research supports examining leniency bias using a small number of profiles per participant. Therefore, using a smaller number of subordinate ratings (such as two subordinates in the *Low Performer* group) is still sufficient for examining leniency bias. Similar methods of calculating leniency bias in the ratings is employed in many studies (i.e., deviation from an average or median rating point (e.g., Bol, 2011; Bol & Smith, 2011).

[38]Both papers base their Social Comparison questions based on the Self-promotion dimension questions from Bolino and Turnley (1999). I also directly reference Bolino and Turnley (1999) when making my adaptations to understand the roots of the questions to ensure my adaptations are appropriate.

of RDG, increasing the rating distribution and reducing leniency bias (i.e., increasing rating accuracy).

Third, I ask nine supplemental questions. The first five supplemental questions explore other potential sources or concerns related to impression management (Appendix E – Panel A – Questions 6, 8, 9, 10, and 11). Two of these additional questions ask about participants' considerations of other parties (their subordinates and superiors). Two ask about their consideration of other factors related to facing a PCC (the extent to which they anticipate they might have to negotiate their ratings within a PCC and the extent to which they believe a PCC might adjust their ratings downward). These supplement impression management questions are intended to assess some of the alternative explanations that might influence impression management concerns and the leniency bias a participant supervisor demonstrates. The last of these first supplement questions asks about impression management from another perspective to provide confirmation of the primary impression management questions.

The second four supplemental questions relate to other potential injunctive norms that may be influenced by the conditions in my study (Appendix E – Panel A – Questions 12-15). These questions relate to factors of distribution, accuracy, and impression management. Two questions ask whether one *should* avoid providing ratings that are too high or too low (a distribution-related norm). One question asks whether one *should* consider the fairness of ratings (an accuracy-related norm). One question asks whether one *should* consider peer supervisors' opinions of their ratings (an impression management-related norm).

I include all nine of these supplemental questions with the primary goal to assess other related factors that may influence rating behaviour. However, they have a secondary beneficial effect of reducing the emphasis on my main social comparison, impression management, and injunctive norm questions to reduce the potential for a demand effect.[39]

4.7 Conclusion

Based on the aforementioned design considerations, I conduct a 2x2 experiment to test my hypotheses regarding how the presence versus absence of a PCC and the presence versus absence of RDG affect supervisors' leniency bias. Participant ratings of ten subordinates are collected to assess the impact of a given scenario on the leniency bias they display. The next chapter discusses the results of this experiment.

[39] Question 7 in the post-experimental questionnaire is an attention check question where participants are asked to provide the answer '2'.

CHAPTER 5: RESULTS

5.1 Introduction

This chapter provides the experiment results. Section 5.2 reports demographic data for my participants and the assessment of potential control variables. Section 5.3 reports testing of the main and interactive effect of PCCs and RDG on leniency bias (Hypothesis 1, 2, and 3) using a hierarchical linear model. Section 0 provides additional testing regarding my theoretical constructs of social comparison, impression management, and injunctive norms. This chapter concludes in Section 5.5.

5.2 Experimental Participants' Demographic Data and Control Variables

5.2.1 Participants Sample

Four hundred Prolific workers participated in this experiment. I remove a total of 53 (13.25%) participants from the sample. This is broken down into forty-seven participants that do not meet the supervisory experience requirements, three participants that provide extreme outlier responses, and three participants with a response pattern suggesting they did not actively engage in the task. The following discussion expands on my reasoning for these exclusions.

First, as discussed in Section 4.3, I use a pre-set screening characteristic on Prolific to recruit participants with supervisory experience. However, since this is a crucial screening variable, I ask a secondary screening question about the supervisory experience within my demographic questions (Appendix E – Panel B – Question 6). Responses to this secondary screening questions included 47 participants who responded with *"0"* to—"*What is the most employees you have supervised at one time"*. This response indicates that they do not, in fact, have supervisory experience since they have not supervised any subordinates. Therefore, I remove these participants from my sample.[40,41]

Second, I drop three participants for providing extreme outlier ratings. I define extreme outliers as being more than three standard deviations from the mean (Aguinis et al., 2013). These participants present as outliers both when I examine an average of the ten *Profile* ratings and are outliers for multiple of their individual *Profile* ratings.

Last, I drop the other three participants who provide the same numeric response across (1) all ten subordinates' profiles, (2) the post-experimental questionnaire, or (3) both. I interpret this response pattern as indicative of non-effortful responses. These three participants also fail the attention check question in the post-experimental questionnaire (Appendix E – Panel A – Question 7). Further, they also complete the study in considerably less than the average time. Therefore, the final sample used in the analysis in this chapter includes 347 participants. [42]

5.2.2 Demographic Information

Table 1 – Panel A provides demographic information on the 347 participants included in my analysis. [43] Participants spend an average of 9.602 minutes (σ = 4.724) on the task. [44] Participants are, on average, 33.784 years old (σ = 8.191; Table 1 – Panel A). They range in age from 20 to 62 years old, with 82.4% of participants falling between ages 25 and 45 (untabulated),

[40] I ***do*** include participants that indicate they have supervised subordinates, but indicate they do not have specific evaluation experience. I include these participants as they still have supervisory experience, and therefore, likely experience guiding, monitoring, and providing feedback to subordinates sufficent to have the skills and background to evaluate subordinate performance. To incorporate any effects of evaluation experience on the subordinate rating outcomes I control for evaluation experience in my analysis.

[41] Inferences regarding hypothesis tests are unaffected by dropping these observations from the analysis.

[42] Inferences from analysis regarding hypothesis testing are unaffected by the inclusion of these six removed observations.

[43] As discussed in Section 4.3, I recruit participants from the online labour platform Prolific pre-screening for supervisory experience, education of at least an undergraduate degree, and living in the United Kingdom, Canada, United States, Australia, or New Zealand

[44] As discussed in Section 4.3, I pay all participants a flat rate of £2, which is equivalent to approximately $3.52 Canadian dollars based on the Bank of Canada Exchange Rate from the first week of March 2021 (the week the study was run; Bank of Canada, n.d.). All participants were paid in British pounds.

36% of participants are male, and 79.3% have evaluation experience (Table 1 – Panel A).[45]

5.2.3 *Leniency Bias*

Based on the *Leniency Bias* measure discussed in Section 0, I sort subordinate profiles into their performance levels (low, moderate, and high) and calculate participants' *Leniency Bias* for each subordinate's profile. A positive outcome of this calculation signifies a participant rated the subordinate higher than the *Subordinate Midpoint Rating*, indicating a lenient rating. This measure is calculated ten times for each participant, once for each of the ten subordinate profiles they rate. Table 1, Panel E provides the descriptive statistics for the *Leniency Bias* measure for each subordinate by condition [e.g., *Leniency_Low_1*]. Next, Table 1, Panel F provides the descriptive statistics of *Leniency Bias* grouped by performance Level.

5.2.4 *Correlation Analysis*

The correlations of demographic information collected with critical independent and dependent variables are shown in Table 2. Of the demographic variables collected, *Evaluation Experience* is significantly correlated with the Average Rating of all the subordinates for the participant (Pearson correlation; -0.090, p = 0.093) and *Low Performers* (Pearson correlation; -0.174, two-tailed p = 0.005; Table 2). Given this significant correlation, I control for evaluation experience (*Evaluation Experience*) in my analysis.

5.3 Test of Hypotheses

5.3.1 *Method of Analysis*

My data structure is such that each participant provides ten subordinate ratings. As such, I

[45] The means of participants' age, gender, and evaluation experience do not significantly vary between conditions (untabulated; all at least p > 0.220, two-tailed).

must use statistical techniques that consider the nested data structure, whereby the ten subordinate ratings are nested under each participant (each participant is assigned an ID number: *ID*). One of the main methods to analyze a nested data structure is using a Hierarchical Linear Model (HLM; Roberts & Fan, 2004). A two-level HLM model allows for the representation of different effects within the nesting criterion (in this case, each participant; also referred to as the Level 2 unit) for different observations (in this case, each *Profile* rating; also referred to as the Level 1 unit; Bryk & Raudenbush, 1988). A distinctive feature of HLM is that regression coefficients are presumed to vary across the Level 2 unit and that the analysis of interest is that variation between the different occurrences of the Level 2 unit (i.e., the participants; Bryk & Raudenbush, 1988, p. 70). More simply, HLM incorporates the unique characteristics of each participant that may influence the 10 *Profile* ratings they assign. HLM controls for these *within* participant variations to allow for better analysis of the difference *between* participants for each rating (Bryk & Raudenbush, 1988). Further, HLM helps to overcome issues with other types of multilevel analysis. Such as aggregation bias, misestimated standard errors, and heterogeneity of regressions (Bryk & Raudenbush, 1988).

To summarize, each participant in my study provides ten responses; as such, it is necessary to control for individual participant-level characteristics. Additionally, I expect each subordinate rating to vary *within* participants and seek to capture the variation of each subordinate rating between participants. HLM allows me to achieve both of these goals. HLM can appropriately control for participant-level characteristics to assess variation between participants. Additionally, HLM allows for analysis of each subordinate *Profile* rating for each participant; without creating issues by aggregating ratings or trying to compare heterogeneous regressions.

The next issue I consider for analysis is that my data violates assumptions of normality. Violations of normality follow ex-ante considerations of the data to be collected. I seek to examine leniency bias, which by definition, would not follow a normal distribution, given that bias is a systematic deviation from the 'accurate' assessment (Saal & Landy, 1977). Non-normality is confirmed ex-post based on a Shapiro-Wilk test (Table 1 – Panel D). Based on this testing, eight of the ten *Leniency Bias* variables are non-normally distributed (for all non-normally distributed *Leniency Bias* variables, W>0.975, p<0.100; see Table 1 – Panel D for individual results). Based on the *a priori* expectation and the *a posteriori* testing, any analysis conducted must be able to handle non-parametric non-normal data.

Given the non-normality of my data, I follow Roberts and Fan (2004) that suggests the use of, and provide instructions for, bootstrapping within an HLM. As per Roberts and Fan (2004, p. 24), bootstrapping within HLM serves two purposes: 1) making non-parametric inferences about parameter estimates and 2) correcting potential bias in parameter estimation. Moreover, Roberts and Fan (2004, p. 24) note that bootstrapping is particularly helpful when data assumptions have been violated, such as data non-normality and when the number of samples within the Level 1 (i.e., subordinate *Leniency Bias*) is small within each Level 2 unit (i.e., participant). As such, combining bootstrapping with HLM helps to overcome both the nested structure of my data (HLM) and the non-normality of my data (bootstrapping). I follow one of the two bootstrapping methods for HLM suggested by Roberts and Fan (2004) and bootstrap with replacement incorporating the nested data structure as that is most appropriate based on my sample.[46]

[46] The other method suggested by Roberts and Fan (2004) is to draw a bootstrapped sample with replacement, but ignoring the nested data structure. Given design features of this study, namely that each subordinate should have a different rating and the nesting by participant, using a bootstrapping method that incorporates the nested design is most appropriate. For robustness, analysis was run using the alternative method of bootstrapping and results do not differ significantly.,

I follow this testing with planned contrast testing following Buckless and Ravenscroft (1990). Buckless and Ravenscroft (1990) suggest using planned contrast coding to test hypotheses that predict an ordinal interaction rather than a disordinal interaction. This method is suggested as it provides a way to compare specific groups or levels of a categorical variable that have a meaningful order or sequence.

Ordinal interactions occur when the effect of one independent variable on the dependent variable changes depending on the level of another independent variable. In other words, the strength or direction of the relationship between the independent variable and the dependent variable depends on the level of the other independent variable. Such as is the case with my third hypothesis. Planned contrast coding allows for testing the specific patterns of differences among the levels of the ordinal variable that are of interest (Buckless & Ravenscroft, 1990). By comparing the means of specific groups or levels of the categorical variable with each other, planned contrast coding allows for a more focused and informative test of the interaction effect.

I use planned contrast coding based on my hypothesized interaction using suggested methods from Buckless & Ravenscroft (1990) and Guggenmos et al. (2018). As per Guggenmos et al. (2018), the exact coding suggested in Buckless and Ravenscroft (1990) may not be adequate for all hypothesis testing, instead custom contrast coding based on predicted hypothesis may be best suited to test hypothesized ordinal interactions. Based on the predicted pattern (depicted in Figure 4), I test two slightly different custom contrast codings. The first is + 1 +2 -2 -1, and the second is +1 +3 -6 +2. These contrast codings represent my predictions: 1) that the control condition will be slightly lenient, 2) that the PCC Only condition will show the most leniency, 3) the RDG Only condition will show the least leniency, and 4) that the PCC and RDG condition will show that the presence of RDG weakens the impact of PCC on leniency bias, thus that the leniency will be less than the PCC only and more than the RDG only.

In summary, I test my hypotheses using a bootstrapped hierarchical linear model. I group ratings by participant (*ID*; Level 2 unit) using the previously created measure of *Leniency Bias* (Level 1 unit; see Section 4.6.1 for creation of the leniency bias measure).

5.3.2 Hypothesis One

Using all ten subordinate profiles, I do not find support for hypothesis one, which predicts a PCC will have a main effect of increasing *Leniency Bias* (*p=0.466*, two-tailed, Table 3 – Panel B). I next partition my sample by *Performance Level.* Within the *Low Performer* group, I find support for my hypothesis one with a significant mean difference (0.056, *p=0.094*, two-tailed; Table 4 - Panel B) between the *PCC* and *No PCC* Conditions.[47] I also find a significant difference among the *High Performers*; however, this significant difference goes against my prediction with a mean difference of -0.069 (*p=0.062*, two-tailed; Table 4 - Panel B). These results seem to indicate that a PCC presence does inflate the leniency bias among *Low Performers* but may decrease the leniency bias among high performers. This finding supports prior research that leniency bias may be asymmetrically applied (Napier & Latham, 1986). As well, given that the effects of a PCC on *Leniency Bias* seem to be differential, the inflation of *Low Performers* and reduction to *High Performers* may actually compress rating and make differentiation more difficult pre-committee, which may affect both the rating discussions in the PCC and lead to subordinates' negative perceptions of fairness in the process.

[47] Statistical inferences of a PCC on Leniency Bias within the *Low Performer* group are robust to a non-bootstrapped three-way mixed ANCOVA (*p = 0.079*, two-tailed; untabulated), run with the same model factors and covariates.

5.3.3 Hypothesis Two

Using all ten subordinate profiles, I find support for hypothesis two, which predicts RDG will have the main effect of reducing *Leniency Bias* (*p=0.027*, two-tailed, Table 3 – Panel B). Further, using my partitioned sample, I find support for hypothesis two within *Low Performers* with a significant mean difference (-0.052, *p=0.027*, two-tailed; Table 4 - Panel B) between the *RDG* and *No RDG* Conditions.[48] I do not find a significant difference neither within *Moderate Performers* (-0.043, *p=0.318*, two-tailed) nor *High Performers* (0.042, *p=0.249*, two-tailed). Overall, these results indicate consistency with prior findings for RDG, with the main effect of reducing bias, and with RDG affecting lower performers more than other performance levels (Bernardin & Villanova, 1986; Friedrich, 1993; Lawler, 1990; Murphy & Cleveland, 1991; e.g., Napier & Latham, 1986).

5.3.4 Hypothesis Three

Using all ten subordinate profiles, I do not find support for hypothesis three, which predicts that RDG will weaken the effects of PCC on *Leniency Bias* (*p=0.825*, two-tailed, Table 3 – Panel B). Additionally, I do not find support for hypothesis three using my partitioned sample: *Low Performers* (*p=0.576*, two-tailed), *Moderate Performers* (*p=0.894*, two-tailed), *High Performers* (*p=0.807*, two-tailed).

To conduct a further test of my hypotheses, I conduct contrast testing based on *a priori* expectations of my pattern of results. Based on the predicted pattern of results, depicted in Figure 4, I test my pattern of results for each *Performance Level*. As shown in Table 5, based on several iterations of the same overall pattern of results, I find support for both the predicted pattern of

[48] Statistical inferences of a RDG on *Leniency Bias* within the *Low Performer* group are robust to a non-bootstrapped three-way mixed ANCOVA (*p = 0.003*, two-tailed; untabulated), run with the same model factors and covariates.

results (*p<0.001*, two-tailed). This pattern predicts that (1) those facing neither a PCC nor RDG will display some leniency bias, (2) those facing a PCC will display the highest leniency bias, (3) those with only RDG will display the least leniency bias, (4) that RDG will weaken the effect of a PCC on leniency bias. Therefore, I find some support for all three hypotheses within the *Low Performer* group based on planned contrast testing.

Based on the planned contrast testing, I did not find significant results for any hypothesis within the *Moderate Performer* or *High Performer* groups (Table 6 and Table 7 respectively; *p>0.232* for all).

5.4 Additional Testing – Theoretical Mechanisms

5.4.1 Principle Components Analysis

To further test my theoretical mechanisms (Figure 3), I look to my post-experimental questions (PEQs) to explore the impression management concerns and the understood norms of participants. I run a principal components analysis (PCA) on post-experimental questions (PEQs) measuring impression management and norms. I assess the suitability of PCA before analysis, and inspection of the correlation matrix shows that all variables had at least one correlation coefficient greater than 0.3.

PCA revealed four components with eigenvalues greater than one, and these explain 20.410%, 20.204%, 13.289%, and 11.957% for a total of 65.860% explanation of the total variance. I employ a Varimax orthogonal rotation to aid interpretability. The rotated solution exhibits a ‘simple structure’ (Thurstone, 1947). Visual inspection of the scree plot indicates that all four components should be retained (Cattell, 1966). In addition, a four-component solution meets the interpretability criterion. As such, I retain all four components.

Overall, the interpretation of the data was consistent with the attributes the questionnaire was designed to measure with strong loadings of Impression Management concerns on Component 1, Distribution Norms on Component 2, Alternative Explanations for rating differences (i.e., fear of negotiation and downward pressure within a PCC) on Component 3, and Fairness and Accuracy Norms on Component 4. Component loadings and communalities of the rotated solution are presented in Table 9. The factor analysis results confirm that the PEQs load into component measures as intended.

5.4.1.1 Impression Management is increased by the presence of a PCC (Hypothesis 1)

Based on the above components, I use Component 1 to examine if the presence of a PCC impacts the impression management [*IM*] concerns of participants (Component 1, Table 9). I calculate an average score of the three questions that make up this *IM* measure.

Univariate ANCOVA testing shows that a PCC has a main effect on *IM* (*p=0.028*, two-tailed) but that RDG does not have a significant direct impact on *IM* concerns (*p=0.369*, two-tailed). Supporting the theoretical development for Hypothesis 1: that the presence of a PCC increases impression management concerns.

5.4.1.2 Injunctive Norms are Strengthened by RDG (Hypothesis 2)

Next, using component 2 from my PCA Factor Analysis [*Dist_Norm*], I examine if RDG increases the strength of the injunctive norms (Component 2, Table 9). To test this, I calculate an average score of the three questions that make up this *Dist_Norm* component. ANCOVA testing on *Dist_Norm* shows that RDG has a significant effect on Dist_Norm (*p=0.092,* two-tailed), but PCC does not (*p=0.341*, two-tailed). [49]

[49] Similar results are obtained using the Fairness and Accuracy Norms factor component.

5.4.1.3 Interactive Effects of PCCs and RDG on Impression Management (Hypothesis 3)

As discussed above, univariate ANCOVA testing shows a PCC has a main effect on *IM* (*p=0.028,* two-tailed), but that RDG does not have a significant direct impact on *IM* concerns (*p=0.369*, two-tailed). However, the concerns a participant has with *Dist_Norm* affect the *IM* concerns a participant has (*p<0.001*, two-tailed). This supports the theoretical links in hypothesis three that a PCC strengthens the impression management concerns, and that RDG while not directly impacting IM does seem to impact the impression management concerns based on the strengthen injunctive norms around the distribution (*Dist_Norm).*

Overall, this lends support that anticipating reporting to a PCC impacts individuals' impression management concerns; additionally, that RDG affects the perceived norms of individuals concerning their perception of the appropriate distribution of ratings. Further that when both a PCC and RDG are used that Impression Management concerns are impacted by both the presence of the PCC and the injunctive norms of the individual.

5.5 Summary

This chapter illustrates the results of the tests of my hypothesis. Generally, I do not find support for the predicted effects of PCCs and RDG on leniency bias when examining a full range of subordinates. However, when reviewing the *Low Performers'* ratings, I find support for the predicted main effects of PCCs and RDG on leniency bias. Thus, I find support for Hypothesis 1 and Hypothesis 2 within this population. I also find some evidence that PCCs work to compress the ratings of subordinates by having a downward effect on the ratings of *High Performers*. I also find support in my planned contrast testing within the *Low Performers* for my overall predicted pattern of results. Further, I find some evidence that a PCC impacts the impression management concerns of individuals and that RDG influences the perceived distribution norms of individuals, supporting my theoretical mechanisms.

CHAPTER 6: CONCLUSION

6.1 Introduction

This chapter discusses a summary of my hypotheses testing results in Section 6.2. Further, in Section 6.3, I identify the limitations of this study and opportunities for future research. Finally, I conclude in Section 6.4.

6.2 Discussion of Hypotheses Testing Results

The results presented in Chapter 5 show support for Hypothesis One and Two among the group of *Low Performers*, and that planned contrast testing shows support for my predicted pattern of results. I also find evidence that a PCC impacts the impression management concerns of individuals and that RDG influences the perceived distribution norms of individuals. As discussed previously, understanding the impact of anticipating a PCC on a supervisor's leniency bias is crucial, as poor-quality information entering the committee may result in poor-quality information leaving the committee.

When analyzing by performance level, I find support for my first and second hypotheses within *Low Performers.* I find support that supervisors may display more leniency in ratings prepared in anticipation of a PCC, especially among low performers. As the increased bias appears to impact low-performers, this may create additional fairness concerns for moderate and high-performers, which could demotivate these subordinates. Additionally, it is essential to be aware of this potential increased leniency bias to ensure effective calibration within the PCC. I find support that RDG does have a main effect of reducing the leniency bias displayed among low performers. Further, using planned contrast testing, I find support for my predicted pattern of results for low performers. That is, the presence of a PCC has the main effect of increasing leniency bias, the presence of RDG has the main effect of reducing leniency bias, and when a PCC is present, the presence of RDG weakens the effect of PCCs on leniency bias. This finding indicates RDG may be helpful in settings with a PCC. However, perhaps a stronger control, such as forced distributions, is required to mitigate more of the impact of PCCs on leniency bias.

6.3 Limitations and Opportunities for Future Research

This study has various limitations that provide opportunities for future research. First, since my participants do not actually participate in a PCC during my experiment, it may be that past experience with a specific PCC may result in learning behaviour; as such future research might examine the impact of multiple rounds of PCCs to explore how PCC experience changes a supervisor's approach over time. Second, characteristics of a subordinate's relationship with a supervisor may create different impacts on the leniency bias displayed towards a particular individual. This study removes that aspect of specific supervisor and subordinate relationships by presenting the same ten anonymized and generic subordinate profiles to supervisors. This design creates a more sterile test of the specific hypothesis. Still, what aspects of relationships impact the supervisor's relationship with that subordinate, or even the nature of the relationships between peer supervisors, may be an interesting aspect to explore in future research. Finally, future research might also examine the effects of different types of calibration committees on leniency bias, as previous papers on Higher-Level Calibration Committees (Demere et al., 2018) and Mixed Calibration Committees (Grabner et al., 2021) find differences in the interactions and outcomes from different types of calibration committees. As such, it is possible that supervisors may prepare performance evaluations in anticipation of these different committees in different ways.

6.4 Conclusions

Overall, my research offers several contributions to practice and theory. First, I contribute to the performance management literature by assessing the leniency bias displayed by supervisors anticipating participating in a PCC. My study is a meaningful extension of studies such as Bol et al. (2019), Grabner et al. (2020), and Demeré et al. (2019) that examine post-committee outcomes in settings where a calibration committee is always present. This study adds an examination of a PCC's impact on supervisors' initial ratings prepared in anticipation of a PCC as compared to those prepared without the expectation of a PCC. Understanding this pre-PCC impact is vital, as the quality of the ratings prepared for a PCC directly impacts the quality of the ratings post-PCC.

Second, I incorporate how RDG interacts with a PCC. In practice, RDG is a common control often used in conjunction with calibration committees (Bol et al., 2019; Demeré et al., 2019; e.g., Mercer, 2013). Given the predicted opposing effects on leniency bias, understanding how the two controls interact is important to inform their joint use in practice. I also provide an extension to prior studies that examine RDG in settings without a PCC (e.g., Stewart et al., 2010).

Third, my study instrument adds the ability to examine additional features. I demonstrate the ability to test social comparison in an online environment; this extends prior literature that examines social comparison in a laboratory environment (Hannan et al., 2013; e.g., Tafkov, 2013). I also offer insights into how impression management, social comparison, and norms are influenced by the presence of a PCC and RDG. Additionally, in contrast to most prior experimental literature that examines leniency bias by examing one or two subordinates, I create an instrument that allows for a comparison of leniency bias between low, moderate, and high performers. By studying a larger group of subordinates, I can better comment on the impact of biases on various levels of subordinate performance, and this greater breadth of performance levels expands the opportunities to test and understand the systematic nature of leniency bias.

REFERENCES

Aguinis, H. (2009). *Performance management*. Pearson Prentice Hall Upper Saddle River, NJ.

Aguinis, H., & Bradley, K. J. (2014). Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods*, *17*(4), 351–371. https://doi.org/10.1177/1094428114547952

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, *16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Albert, L. (2017, June 16). *More Companies Using Calibration to Assess Talent*. Talent Management & HR. https://www.tlnt.com/more-companies-using-calibration-to-assess-talent/

Aronson, E., Wilson, T. D., & Akert, R. M. (2010). *Social Psychology* (7th ed.). Prentice Hal.

Arshad, F. (2020). *Performance Management Systems in Modern Organizations*. https://doi.org/10.26116/CENTER-LIS-2003

Arshad, F., Cardineal, E., & Dierynck, B. (2020). *Facing a Calibration Committee: The Impact on Costly Information Collection and Subjective Performance Evaluation*. Working Paper.

Bailey, W. J., Hecht, G., & Towry, K. L. (2011). Dividing the Pie: The Influence of Managerial Discretion Extent on Bonus Pool Allocation. *Contemporary Accounting Research*, *28*(5), 1562–1584. https://doi.org/10.1111/j.1911-3846.2011.01073.x

Baker, G. P., Jensen, M. C., & Murphy, K. J. (1988). Compensation and Incentives: Practice vs. Theory. *The Journal of Finance*, *43*(3), 593–616. https://doi.org/10.2307/2328185

Bank of Canada. (n.d.). *Daily exchange rates* [Government]. Bank of Canada. Retrieved February 6, 2023, from https://www.bankofcanada.ca/rates/exchange/daily-exchange-rates/

Barrett, J. (1966). The conflict of generations in college counseling. *Mental Hygiene*, *50*(1), 111–116.

Bates, S. (2003). Forced Rankling. *HR Magazine*, *48*, 62–68.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Beach, S., & Tesser, A. (1995). Self-Esteem and the Extended Self-Evaluation Maintenance Model. *Self-Esteem and the Extended Self-Evaluation Maintenance Model: The Self in Social Context*. https://doi.org/10.1007/978-1-4899-1280-0_8

Beer, M., & Gery, G. J. (1972). Individual and organizational correlates of pay system preferences. *Managerial Motivation and Compensation*, 325–349.

Bénabou, R., & Tirole, J. (2005). Self=Confidence and Personal Motivation. *Psychology, Rationality and Economic Behaviour: Challenging Standard Assumption*, 19–57.

Bentley, J. W. (2019). Decreasing Operational Distortion and Surrogation Through Narrative Reporting. *The Accounting Review*, *94*(3), 27–55. https://doi.org/10.2308/accr-52277

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at work*. Boston, Ma.: Kent Publishing Company.

Bernardin, H. J., & Villanova, P. (1986). *Generalizing from laboratory to field settings*. Lexington: Lexington Books.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Blume, B. D., Baldwin, T. T., & Rubin, R. S. (2009). Reactions to Different Types of Forced Distribution Performance Evaluation Systems. *Journal of Business and Psychology*, *24*(1), 77–91. https://doi.org/10.1007/s10869-009-9093-5

Blume, B., Rubin, R., & Baldwin, T. (2013). Who is Attracted to an Organization Using a Forced Distribution Performance Management System? *Human Resource Management Journal*, *23*, 360–378. https://doi.org/10.1111/1748-8583.12016

Bol, J. C. (2008). Subjectivity in compensation contracting. *Journal of Accounting Literature, 27*, 1–24.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review*, *86*(5), 1549–1575.

Bol, J. C., Aguiar, A. B., & Lill, J. B. (2019, August 27). *Peer-Level Calibration of Performance Evaluation Ratings: Are There Winners or Losers?* https://www.semanticscholar.org/paper/Peer-Level-Calibration-of-Performance-Evaluation-or-Bol-Aguiar/9100035b67e4ac3a93e80b36947e1784fd7b1b5a

Bol, J. C., & Smith, S. D. (2011). Spillover Effects in Subjective Performance Evaluation: Bias and the Asymmetric Influence of Controllability. *The Accounting Review*, *86*(4), 1213–1230. https://doi.org/10.2308/accr-10038

Bolino, M. C., & Turnley, W. H. (1999). Measuring impression management in organizations: A scale development based on the Jones and Pittman taxonomy. *Organizational Research Methods*, *2*(2), 187–206.

Bolt, D. M., & Johnson, T. R. (2009). Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Bonner, S. (2008). *Judgment and decision making in accounting*. Pearson/Prentice Hall.

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management*, *18*(2), 321–352. https://doi.org/10.1177/014920639201800206

Brown, D., Ferris, L., Heller, D., & Keeping, L. (2007). Antecedents and consequences of the frequency of upward and downward social comparison at work. *Organizational Behavior and Human Decision Processes*, *102*, 59–75.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model. *American Journal of Education*, *97*(1), 65–108. http://www.jstor.org/stable/1084940

Buchheit, S., Dalton, D. W., Pollard, T. J., & Stinson, S. R. (2019). Crowdsourcing Intelligent Research Participants: A Student versus MTurk Comparison. *Behavioral Research in Accounting*, *31*(2), 93–106. https://doi.org/10.2308/bria-52340

Buchheit, S., Doxey, M. M., Pollard, T., & Stinson, S. R. (2018). A Technical Guide to Using Amazon's Mechanical Turk in Behavioral Accounting Research. *Behavioral Research in Accounting*, *30*(1), 111–122. https://doi.org/10.2308/bria-51977

Buckless, F. A., & Ravenscroft, S. P. (1990). Contrast Coding: A Refinement of ANOVA in Behavioral Analysis. *The Accounting Review*, *65*(4), 933–945. https://www.jstor.org/stable/247659

Callan, M. J., Kim, H., Gheorghiu, A. I., & Matthews, W. J. (2017). The Interrelations Between Social Class, Personal Relative Deprivation, and Prosociality. *Social Psychological and Personality Science*, *8*(6), 660–669. https://doi.org/10.1177/1948550616673877

Caruso, K. N. (2013). *A practical guide to performance calibration: A step-by-step guide to increasing the fairness and accuracy of performance appraisal.* http://cdn2.hubspot.net/hub/91252/file-338193964-pdf/Practical_Guide_to_Performance_Calibration_October_2013.pdf

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology* (Vol. 24, pp. 201–234). Elsevier.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity, and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (pp. 151–192). Oxford University Press.

Cohen-Charash, Y., & Spector, P. E. (2001). The role of justice in organizations: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *86*(2), 278–321.

Colquitt, J. A., & Chertkoff, J. M. (2002). Explaining injustice: The interactive effect of explanation and outcome on fairness perceptions and task motivation. *Journal of Management*, *28*(5), 591–610.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425.

de Araújo, F. F. (2014). Do I Look Good In Green?: A Conceptual Framework Integrating Employee Green Behavior, Impression Management, and Social Norms. *Amazônia, Organizações e Sustentabilidade*, *3*(2), 7–23. https://doi.org/10.17800/2238-8893/aos.v3n2p7-23

Demeré, B. W., Sedatole, K. L., & Woods, A. (2019). The Role of Calibration Committees in Subjective Performance Evaluation Systems. *Management Science*, *65*(4), 1562–1585. https://doi.org/10.1287/mnsc.2017.3025

Dineen, B. R., Lewicki, R. J., & Tomlinson, E. C. (2006). Supervisory guidance and behavioral integrity: Relationships with employee citizenship and deviant behavior. *Journal of Applied Psychology*, *91*(3), 622.

Dittmar, H., Bond, R., Hurst, M., & Kasser, T. (2014). The relationship between materialism and personal well-being: A meta-analysis. *Journal of Personality and Social Psychology*, *107*(5), 879.

Endler, N. S. (1993). *Personality: An interactional perspective*. Springer.

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

Ewenstein, B., Hancock, B., & Komm, A. (2016). Ahead of the curve: The future of performance management. *The McKinsey Quarterly*, 1–10. https://www.proquest.com/docview/2699769042/abstract/37EE3F1207D04CD2PQ/1

Fang, H., & Moscarini, G. (2002). Overconfidence, morale and wage-setting policies. In *Cowles Foundation Discussion Paper* (p. 1422).

Farrell, A. M., Grenier, J. H., & Leiby, J. (2017). Scoundrels or Stars? Theory and Evidence on the Quality of Workers in Online Labor Markets. *The Accounting Review*, *92*(1), 93–114. https://doi.org/10.2308/accr-51447

Farrell, M., & Sweeney, B. (2021). Amazon's MTurk: A currently underutilised resource for survey researchers? *Accounting, Finance, & Governance Review*, *27*(1), 36–53.

Ferris, G. R., & Judge, T. A. (1991). Personnel/Human Resources Management: A Political Influence Perspective. *Journal of Management*, *17*(2), 447–488. https://doi.org/10.1177/014920639101700208

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Friedrich, J. (1993). Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review, 100*(2), 298.

Garcia, S. M., & Tor, A. (2007). Rankings, standards, and competition: Task vs. scale comparisons. *Organizational Behavior and Human Decision Processes*, *102*(1), 95–108. https://doi.org/10.1016/j.obhdp.2006.10.004

Garrow, L. A., Chen, Z., Ilbeigi, M., & Lurkin, V. (2020). A new twist on the gig economy: Conducting surveys on Amazon Mechanical Turk. *Transportation*, *47*, 23–42.

Gibbs, M., Merchant, K. A., Van der Stede, W. A., & Vargus, M. E. (2004). Determinants and Effects of Subjectivity in Incentives. *The Accounting Review*, *79*(2), 409–436. https://www.jstor.org/stable/3203250

Goethals, G. R., & Darley, J. M. (1977). Social comparison theory: An attributional approach. *Social Comparison Processes: Theoretical and Empirical Perspectives*, 259–278.

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Doubleday.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

Grabner, I., Künneke, J., & Moers, F. (2020). How Calibration Committees Can Mitigate Performance Evaluation Bias: An Analysis of Implicit Incentives. *The Accounting Review*, *95*(6), 213–233. https://doi.org/10.2308/tar-2016-0662

Greenberg, J. (1990). Organizational justice: Yesterday, today, and tomorrow. *Journal of Management*, *16*(2), 399–432.

Grote, R. (2005). *Forced ranking: Making performance management work*. Harvard Business School Press.

Grund, C., & Przemeck, J. (2012). Subjective performance appraisal and inequality aversion. *Applied Economics*, *44*(17), 2149–2155. https://doi.org/10.1080/00036846.2011.560109

Guggenmos, R. D., Piercey, M. D., & Agoglia, C. P. (2018). Custom Contrast Testing: Current Trends and a New Approach. *Accounting Review*, *93*(5), 223–244. https://doi.org/10.2308/accr-52005

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

Harkins, S., & Jackson, J. (1985). The role of evaluation in eliminating social loafing. *Personality and Social Psychology Bulletin*, *11*, 457–465.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management*, *20*(4), 737–756.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, *41*(1), 43–62.

Hastings, R. (2011). *Survey: Most large firms calibrate performance*. Society for Human Resource Management (SHRM. https://www.shrm.org/ResourcesAndTools/hr-topics/employee-relations/Pages/CalibratePerformance.aspx

Hecht, G., Hobson, J., & Wang, L. (2020). The Effect of Performance Reporting Frequency on Employee Performance. *The Accounting Review*, *95*(4), 199–218.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology*, *63*(5), 579–588. https://doi.org/10.1037/0021-9010.63.5.579

Hughes, R., & Huby, M. (2002). The application of vignettes in social and nursing research. *Journal of Advanced Nursing*, *37*(4), 382–386. https://doi.org/10.1046/j.1365-2648.2002.02100.x

Hull, C. L. (1928). *Aptitude testing.*

Ilgen, D. R., Mitchell, T. R., & Fredrickson, J. W. (1981). Poor performers: Supervisors' and subordinates' responses. *Organizational Behavior and Human Performance*, *27*(3), 386–410.

Iyengar, R., Van den Bulte, C., & Valente, T. W. (2011). Opinion Leadership and Social Contagion in New Product Diffusion. *Marketing Science*, *30*(2), 195–212. https://doi.org/10.1287/mksc.1100.0566

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, *50*(4), 905–925.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychological Perspectives on the Self*, *1*(1), 231–262.

Kaptein, M. (2015). The effectiveness of ethics programs: The role of scope, composition, and sequence. *Journal of Business Ethics*, *132*, 415–431.

Klein, W. M., & Kunda, Z. (1992). Motivated person perception: Constructing justifications for desired beliefs. *Journal of Experimental Social Psychology*, *28*(2), 145–168.

Krupnikov, Y., & Levine, A. S. (2014). Cross-Sample Comparisons and External Validity. *Journal of Experimental Political Science*, *1*(1), 59–80. https://doi.org/10.1017/xps.2014.7

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72.

Lawler, E. E. (1990). *Strategic pay: Aligning organizational strategies and pay systems.* Jossey-Bass.

Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, *46*(4), 352.

Leary, M., & Kowalski, R. (1990). Impression management: A literature-review and two-component model. *Psychological Bulletin*, *107*(1), 34–47.

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

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

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

Libby, T., Salterio, S. E., & Webb, A. (2004). The Balanced Scorecard: The Effects of Assurance and Process Accountability on Managerial Judgment. *The Accounting Review*, *79*(4), 1075–1094.

Lillis, A. M., Malina, M. A., & Mundy, J. (2017). *Rendering Subjectivity Informative in Performance Measurement and Reward Systems: Field Study Insights*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2998471

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the Mask: The Politics of Employee Appraisal. *The Academy of Management Executive (1987-1989)*, *1*(3), 183–193. https://www.jstor.org/stable/4164751

Marreiros, H., Tonin, M., Vlassopoulos, M., & Schraefel, M. C. (2017). Now that you mention it”: A survey experiment on information, inattention and online privacy. *Journal of Economic Behavior & Organization*, *140*, 1–17.

McBriarty, M. A. (1988). Performance appraisal: Some unintended consequences. *Public Personnel Management*, *17*(4), 421–434.

McKinney, J. A., Emerson, T. L., & Neubert, M. J. (2010). The effects of ethical codes on ethical perceptions of actions toward stakeholders. *Journal of Business Ethics*, *97*, 505–516.

Mercer. (2013). *Global performance management survey report (GPMSR).* https://www.mercer.com/content/dam/mercer/attachments/global/Talent/Assess-BrochurePerfMgmt.pdf

Mercer. (2019). *Mercer's 2019 Global Performance Management Survey: Executive Summary*. https://www.imercer.com/uploads/common/HTML/LandingPages/AnalyticalHub/june2019-mercer-2019-global-performance-management-survey-executive-summary.pdf

Merchant, K., & Van der Stede, W. (2017). *Management control systems: Performance measurement, evaluation and incentives* (4th ed.). Pearson Education Limited.

Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a Performance Appraisal Context: The Effect of Audience and Form of Accounting on Rater Response and Behavior. *Journal of Management*, *33*(2), 223–252. https://doi.org/10.1177/0149206306297633

Meyer, H. H. (1975). The pay-for-performance dilemma. *Organizational Dynamics*, *3*(3), 39–50.

Milkovich, G. T., & Newman, J. M. (1993). *Compensation*. Irwin. https://books.google.ca/books?id=4ERYAAAAYAAJ

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

Morris, M. W., Hong, Y., Chiu, C., & Liu, Z. (2015). Normology: Integrating insights about social norms to understand cultural dynamics. *Organizational Behavior and Human Decision Processes, 129,* 1–13.

Murphy, K. R., & Cleveland, J. N. (1991). *Performance appraisal: An organizational perspective* (pp. xiv, 349). Allyn & Bacon.

Napier, N. K., & Latham, G. P. (1986). Outcome expectancies of people who conduct performance appraisals. *Personnel Psychology*, *39*, 827–837. https://doi.org/10.1111/j.1744-6570.1986.tb00597.x

O'Boyle Jr, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, *65*(1), 79–119.

Olson, C., & Davis, G. M. (2003). Pros and cons of forced ranking and other relative performance ranking systems. *Society for Human Resource Management Legal Report*. http://www.shrm.org/hrresources/lrpt_published

Orne, M. T. (1962). Hypnotically induced hallucinations. In *Hallucinations* (pp. 211–219). Grune & Stratton.

Osborne, T., & McCann, L. A. (2004). Forced ranking and age-related employment discrimination. *Hum. Rts.*, *31*, 6.

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Paulhus, D. L. (1991). *Measurement and control of response bias.*

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Pelfrey, S., & Peacock, E. (1991). Ethical Codes of Conduct Are Improving. *Business Forum*, *16*(2), 14–17.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*(1), 7–63.

Prendergast, C., & Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review*, *27,* 355–365.

Prolific. (n.d.). Prolific. https://prolific.ac

Reno, R. R., Cialdini, R. B., & Kallgren, C. A. (1993). The transsituational influence of social norms. *Journal of Personality and Social Psychology*, *64*(1), 104.

Risher, H. (2011). Getting performance management on track. *Compensation & Benefits Review*, *43*(5), 273–281.

Risher, H. (2014). Reward management depends increasingly on procedural justice. *Compensation & Benefits Review*, *46*(3), 135–138.

Robert B. Cialdini, Kallgren, C. A., & Reno, R. R. (1991). A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. In *Advances in Experimental Social Psychology* (Vol. 24, pp. 201–234). Elsevier. https://doi.org/10.1016/S0065-2601(08)60330-5

Roberts, J. K., & Fan, X. (2004). Bootstrapping within the Multilevel/Hierarchical Linear Modeling Framework: A Primer for Use with SAS and SPLUS. *Multiple Linear Regression Viewpoints*, *30*(1), 23–34.

Rosaz, J., & Villeval, M. C. (2012). Lies and biased evaluation: A real-effort experiment. *Journal of Economic Behavior and Organization*, *84*(2), 537–549.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413.

Saal, F. E., & Landy, F. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance, 18,* 19–35.

Sammer, J. (2008). *Calibrating consistency. Society for Human Resource Management (SHRM*. https://www.shrm.org/hr-today/news/hr-magazine/pages/1hr%20management%20agenda.aspx

Schleicher, D. J., Bull, R. A., & Green, S. G. (2009). Rater reactions to forced distribution rating systems. *Journal of Management*, *35*(4), 899–927.

Schlenker, B. R. (1980). *Impression Management: The Self-Concept, Social Identity, and Interpersonal Relations*. Brooks/Cole.

Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology*, *68*(3), 407.

Schneider, D. J. (1981). Tactical self-presentations: Toward a broader conception. In *Impression management theory and social psychological research* (pp. 23–40). Academic Press.

Schneier, C. E. (1977). Operational utility and psychometric characteristics of Behavioral Expectation Scales: A cognitive reinterpretation. *Journal of Applied Psychology, 62*(5), 541–548. https://doi.org/10.1037/0021-9010.62.5.541

Schrage, M. (2000). How the bell curve cheats you. *Fortune*, *141*, 296. https://archive.fortune.com/magazines/fortune/fortune_archive/2000/02/21/273841/index.htm

Schwartz, S. H. (1977). Normative influences on altruism. *Advances in experimental social psychology*, *10*(1), 221–279.

Scullen, S., Bergey, P., & Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A Baseline Simulation. *Personnel Psychology*, *58*, 1–32.

Sears, D. O. (1986). College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature. *Journal of Personality and Social Pyschology*, *51*(3), 515–530.

Shoemaker, N., Curtis, M. B., Fayard, L., & Kelly, M. (2020). What happens when formal and informal norms conflict for IT usage? *Journal of Information Systems*, *34*(2), 235–256.

Shore, L. M., & Thornton, G. C. (1986). Effects of Gender on Self- and Supervisory Ratings. *The Academy of Management Journal*, *29*(1), 115–129. https://doi.org/10.2307/255863

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review*, *11*(1), 22–40.

Smith, R. H. (2000). Assimilative and contrastive emotional reactions to upward and downward social comparisons. *Handbook of Social Comparison: Theory and Research*, 173–200.

Spence, J. R., & Keeping, L. (2011). Conscious rating distortion in performance appraisal: A review, commentary, and proposed framework for research. *Human Resource Management Review*, *21*(2), 85–95.

Stewart, S. M., Gruys, M. L., & Storm, M. (2010). Forced distribution performance evaluation systems: Advantages, disadvantages and keys to implementation. *Journal of Management & Organization*, *16*(1), 168–179.

Suls, J., & Wheeler, L. (2000). A selective history of classic and neo-social comparison theory. *Handbook of Social Comparison: Theory and Research*, 3–19.

Tafkov, I. D. (2013). Private and public relative performance information under different compensation contracts. *The Accounting Review*, *88*(1), 327–350.

Tayler, W. B., & Bloomfield, R. J. (2011). Norms, conformity, and controls. *Journal of Accounting Research*, *49*(3), 753–790.

Taylor, S. (2006). Acquaintance, meritocracy and critical realism: Researching recruitment and selection processes in smaller and growth organizations. *Human Resource Management Review*, *16*(4), 478–489. https://doi.org/10.1016/j.hrmr.2006.08.005

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. *Advances in Experimental Social Psychology*, *21*, 181–228.

Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. In *Advances in experimental social psychology* (Vol. 25, pp. 331–376). Elsevier.

Thurstone, L. L. (1947). The calibration of test items. *American Psychologist*, *2*(3), 103.

Tiffin, J. (1947). *Industrial merit rating.*

Trevino, L. K., & Nelson, K. (2014). *Managing Business Ethics – Straight Talk About How To Do It Right* (6th ed.). Wiley.

von Sydow, M., Braus, N., & Hahn, U. (2019). On the ignorance of group-level effects: The tragedy of personnel evaluation? *Journal of Experimental Psychology: Applied*, *25*(3), 491–515.

Wason, K. D., Polonsky, M. J., & Hyman, M. R. (2002). Designing Vignette Studies in Marketing. *Australasian Marketing Journal*, *10*(3), 41–58. https://doi.org/10.1016/S1441-3582(02)70157-2

Webb, A., Jeffrey, S. A., & Schulz, A. (2010). Factors affecting goal difficulty and performance when employees select their own performance goals: Evidence from the field. *Journal of Management Accounting Research*, *22*(1), 209–232.

West, B. J., Deering, B., & Deering, W. D. (1995). *The lure of modern science: Fractal thinking* (Vol. 3). World Scientific.

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response Biases. In *The ITC International Handbook of Testing and Assessment* (pp. 349–363). Oxford University Press. http://kar.kent.ac.uk/49093/

Wilkie, D. (2015, August 19). *Is the Annual Performance Review Dead?* SHRM. https://www.shrm.org/resourcesandtools/hr-topics/employee-relations/pages/performance-reviews-are-dead.aspx

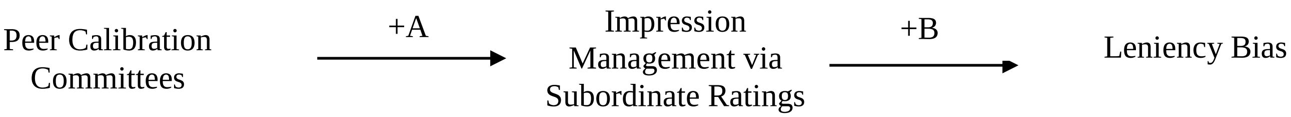
Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*(3), 189–205.

Yukl, G., & Falbe, C. (1990). Influence tactics and objectives in upward, downward, and lateral influence attempts. *The Journal of Applied Psychology*, *75*(2), 132–140.

Zizzo, D. J. (2010). Experimenter Demand Effects in Economic Experiments. *Experimental Economics*, *13*, 75–98. https://doi.org/10.1007/s10683-009-0230-z

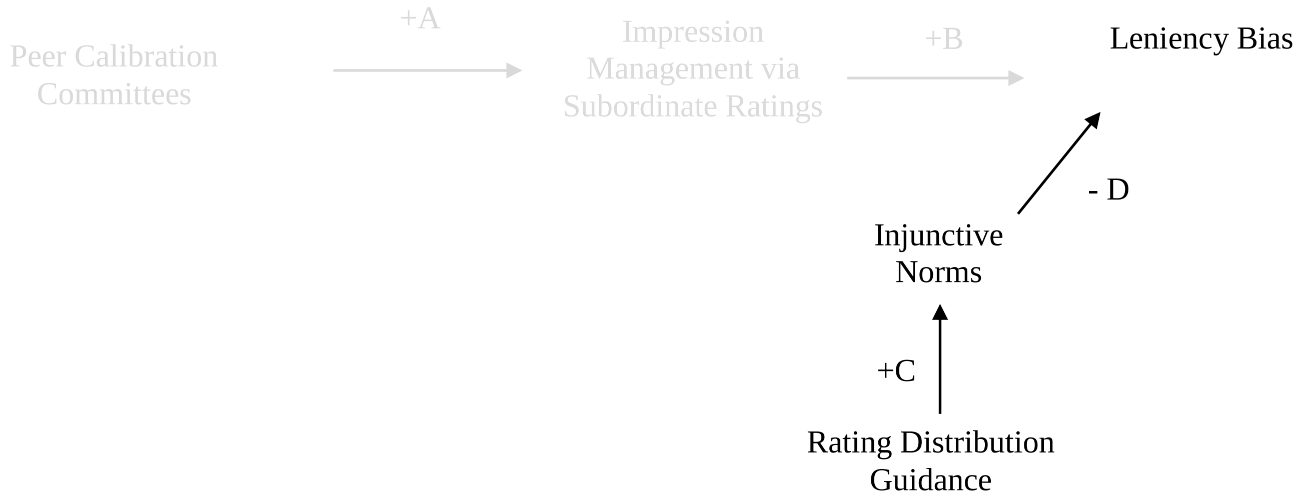
FIGURES

Figure 1 - Hypothesis One


Peer Calibration Committees
+A
Impression Management via Subordinate Ratings
+B
Leniency Bias

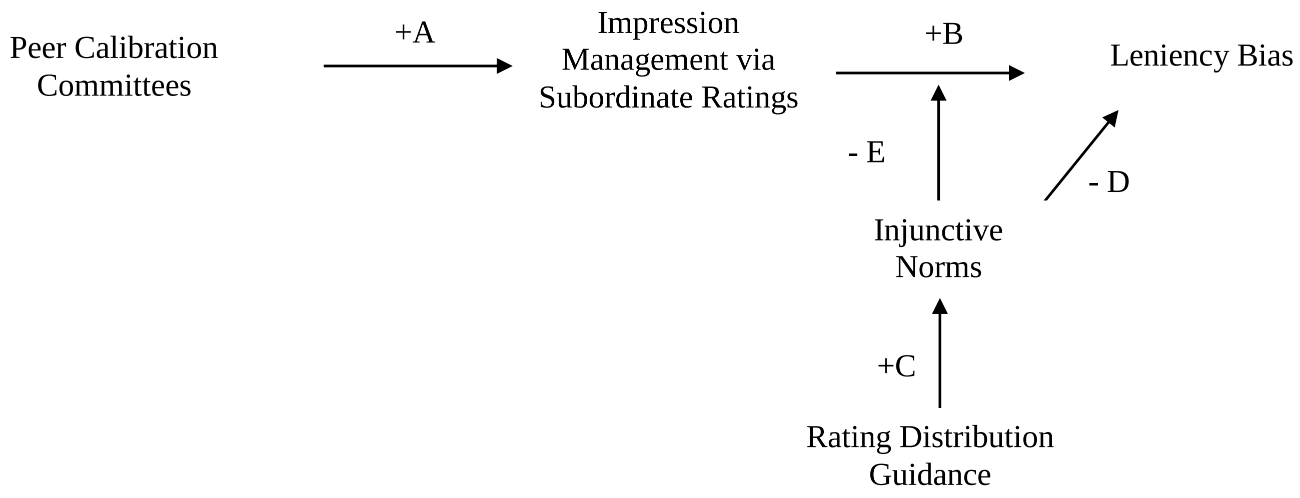
Note. This figure represents the predictions of Hypothesis One. That is, Peer Calibration Committees increase the desire to Manage Impressions (Link A) through higher subordinate ratings to present themselves as better-performing managers. Thereby, the induced impression management concerns increase Leniency Bias (Link B) in subordinate ratings.

Figure 2 - Hypothesis Two


+A
Peer Calibration Committees
Impression Management via Subordinate Ratings
+B
Leniency Bias
- D
Injunctive Norms
+C
Rating Distribution Guidance

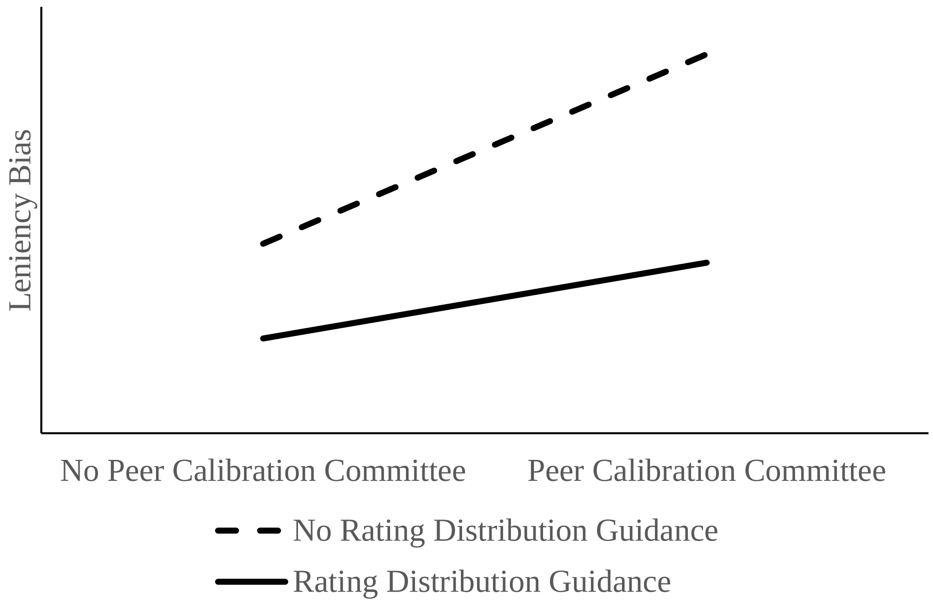
Note. This figure represents the predictions of Hypothesis Two. That is, RDG increases the saliency and the accountability to the injunctive norm for accuracy in a performance evaluation rating (Link C). Further, the increased saliency and accountability to the injunctive norm for accuracy will lead to a reduction in leniency bias (Link D).

Figure 3 - Hypothesized Conceptual Model


Peer Calibration Committees
+A
Impression Management via Subordinate Ratings
+B
Leniency Bias
- E
- D
Injunctive Norms
+C
Rating Distribution Guidance

Note. This figure represents the predictions of Hypothesis Three. When a Peer Calibration Committee is present, RDG increases the saliency of the injunctive norm for accuracy (Link C), which in conjunctive with the increased accountability imposed by the PCC weakens (moderates) the effect of impression management on leniency bias (Link E).

Figure 4 - Graphical Depiction of Hypotheses



Note. This figure represents a graphical depiction of the three hypotheses. That PCC has a main effect of increasing leniency bias (H1), that RDG has a main effect of reducing leniency bias (H2), and that the impact of a PCC is weaker in the presence of RDG (H3).

TABLES

Table 1 – Descriptive and Summary Statistics

Panel A - Demographic Information of Participants

	Control	PCC Only	RDG Only	PCC and RDG	Total
Number of Participants	85	85	91	86	347
Average Age	33.624	33.235	33.231	35.070	33.784
Percentage of Female Participants	62.353%	64.706%	62.637%	66.279%	63.977%
Participants with Evaluation Experience	78.824%	72.941%	85.714%	79.070%	79.251%
Average Number of Subordinates Supervised	13.565	6.929	8.813	10.988	10.055
Average Time to Complete Study (Minutes)	8.844	9.642	9.303	10.625	9.602

Note. The above summary statistics are based on the demographic information collected from participants. Statistics are shown by condition and as the total for all conditions.

Panel B – Summary Statistics of Each Subordinate Rating Raw Score by Condition

	Control			**PCC Only**			**RDG Only**			**PCC and RDG**			**All Conditions**		
	Median	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**
Low Performers															
Low_Performer_1	3.000	3.141	0.560	3.000	3.247	0.486	3.000	3.011	0.527	3.000	3.070	0.504	3.000	3.115	0.525
Low_Performer_2	4.000	4.094	0.526	4.000	4.188	0.422	4.000	3.923	0.500	4.000	4.047	0.612	4.000	4.061	0.526
Low_Performer_3	4.000	4.400	0.539	4.000	4.447	0.523	4.000	4.253	0.589	4.000	4.209	0.511	4.000	4.326	0.549
Moderate Performers															
Moderate_Performer_1	7.000	6.753	0.671	7.000	6.788	0.599	7.000	6.780	0.554	7.000	6.779	0.495	7.000	6.775	0.580
Moderate_Performer_2	6.000	5.541	0.646	5.000	5.435	0.566	5.000	5.286	0.735	5.000	5.267	0.602	5.000	5.380	0.649
Moderate_Performer_3	5.000	5.435	0.626	5.000	5.329	0.625	5.000	5.473	0.621	5.000	5.349	0.589	5.000	5.398	0.616
Moderate_Performer_4	6.000	6.471	0.547	7.000	6.553	0.546	6.000	6.495	0.545	6.000	6.500	0.569	7.000	6.504	0.550
High Performers															
High_Performer_1	8.000	7.788	0.537	8.000	7.718	0.526	8.000	7.758	0.479	8.000	7.744	0.578	8.000	7.752	0.529
High_Performer_2	7.000	6.800	0.573	7.000	6.776	0.564	7.000	6.868	0.581	7.000	6.767	0.588	7.000	6.804	0.576
High_Performer_3	8.000	8.176	0.581	8.000	8.035	0.522	8.000	8.242	0.565	8.000	8.174	0.598	8.000	8.159	0.570
N			**85**			**85**			**91**			**86**			**347**

Note. The above summary statistics are based on the subordinate raw rating scores sorted by performance level. Statistics are shown by performance level and as the total for all conditions.

Panel C – Summary Statistics of Performance Level Group Mean of Subordinate Rating Raw Scores by Experimental Condition

		N	Mean	Std Dev	Minimum	25th Percentile	Median	75th Percentile	Maximum
All Subordinates	Control	85	5.860	1.657	2.000	4.000	6.000	7.000	9.000
	PCC Only	85	5.852	1.593	2.000	5.000	6.000	7.000	9.000
	RDG Only	91	5.809	1.724	1.000	4.000	6.000	7.000	9.000
	PCC and RDG	86	5.791	1.690	2.000	4.000	6.000	7.000	10.000
	Total across conditions	347	5.827	1.67	1.000	4.000	6.000	7.000	10.000
Low Performers	Control	85	3.878	0.442	2.667	3.667	3.667	4.000	5.000
	PCC Only	85	3.961	0.403	3.333	3.667	3.667	4.333	5.000
	RDG Only	91	3.729	0.415	2.333	3.667	3.667	4.000	4.667
	PCC and RDG	86	3.775	0.417	2.000	3.667	3.667	4.000	4.667
	Total across conditions	347	3.833	0.744	1.000	3.000	4.000	4.000	6.000
Moderate Performers	Control	85	6.050	0.415	5.000	5.750	6.250	6.250	7.000
	PCC Only	85	6.026	0.372	5.000	5.750	6.000	6.250	7.000
	RDG Only	91	6.008	0.400	4.750	5.750	6.000	6.250	7.000
	PCC and RDG	86	5.974	0.316	5.250	5.750	6.000	6.250	6.500
	Total across conditions	347	6.014	0.872	3.000	5.000	6.000	7.000	8.000
High Performers	Control	85	7.588	0.429	6.667	7.333	7.667	8.000	8.667
	PCC Only	85	7.510	0.332	6.667	7.333	7.333	7.667	8.333
	RDG Only	91	7.623	0.386	6.333	7.333	7.667	8.000	8.667
	PCC and RDG	86	7.562	0.422	6.667	7.333	7.667	7.667	9.000
	Total across conditions	347	7.572	0.796	5.000	7.000	8.000	8.000	10.000

Note. The above summary statistics are based on the average of the raw subordinate rating scores for each performance level. Statistics are shown by performance level and as the total for all subordinates.

Panel D – Summary Statistics of the Leniency Bias Rating for Each Subordinate by Condition

	Control		PCC Only		RDG Only		PCC and RDG		All Conditions				
											Shapiro-Wilk[a]		
	Mean	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**W**	**p-value[b]**	
Leniency_Low_1	0.141	0.560	0.247	0.486	0.011	0.527	0.070	0.504	0.115	0.525	0.985	<0.001	***
Leniency_Low_2	0.094	0.526	0.188	0.422	-0.077	0.500	0.047	0.612	0.061	0.526	0.997	0.836	
Leniency_Low_3	0.400	0.539	0.447	0.523	0.253	0.589	0.209	0.511	0.326	0.549	0.990	0.011	**
Leniency_Moderate_1	0.253	0.671	0.288	0.599	0.280	0.554	0.279	0.495	0.275	0.580	0.975	<0.001	***
Leniency_Moderate_2	0.041	0.646	-0.065	0.566	-0.214	0.735	-0.233	0.602	-0.120	0.649	0.993	0.084	*
Leniency_Moderate_3	-1.065	0.626	-1.171	0.625	-1.027	0.621	-1.151	0.589	-1.102	0.616	0.993	0.113	
Leniency_Moderate_4	-0.029	0.547	0.053	0.546	-0.005	0.545	0.000	0.569	0.004	0.550	0.990	0.015	***
Leniency_High_1	0.288	0.537	0.218	0.526	0.258	0.479	0.244	0.578	0.252	0.529	0.988	0.007	**
Leniency_High_2	-0.200	0.573	-0.224	0.564	-0.132	0.581	-0.233	0.588	-0.196	0.576	0.992	0.072	*
Leniency_High_3	0.176	0.581	0.035	0.522	0.242	0.565	0.174	0.598	0.159	0.570	0.992	0.072	*
N		**85**		**85**		**91**		**86**		**347**			

Note. In calculating the Leniency Bias measure, the median of the midpoints is taken as the non-biased ratings of that subordinate profile (*Subordinate Midpoint Rating*). Therefore, to evaluate the bias displayed, this *Subordinate Midpoint Rating* is subtracted from the rating given by the participants to create a measure of bias shown by that participant. (See Table 1 - Panel D for the midpoints of each subordinate profile). Statistics are sorted by performance level, shown by both conditions and as a total.

[a] To test the normality of data, a Shapiro-Wilk calculation is conducted such that a significant Shapiro-Wilk indicates that the data is non-normally distributed. Displayed are the Shapiro-Wilk results based on the distribution for all conditions.

[b] Significance is determined based on two-tailed testing.

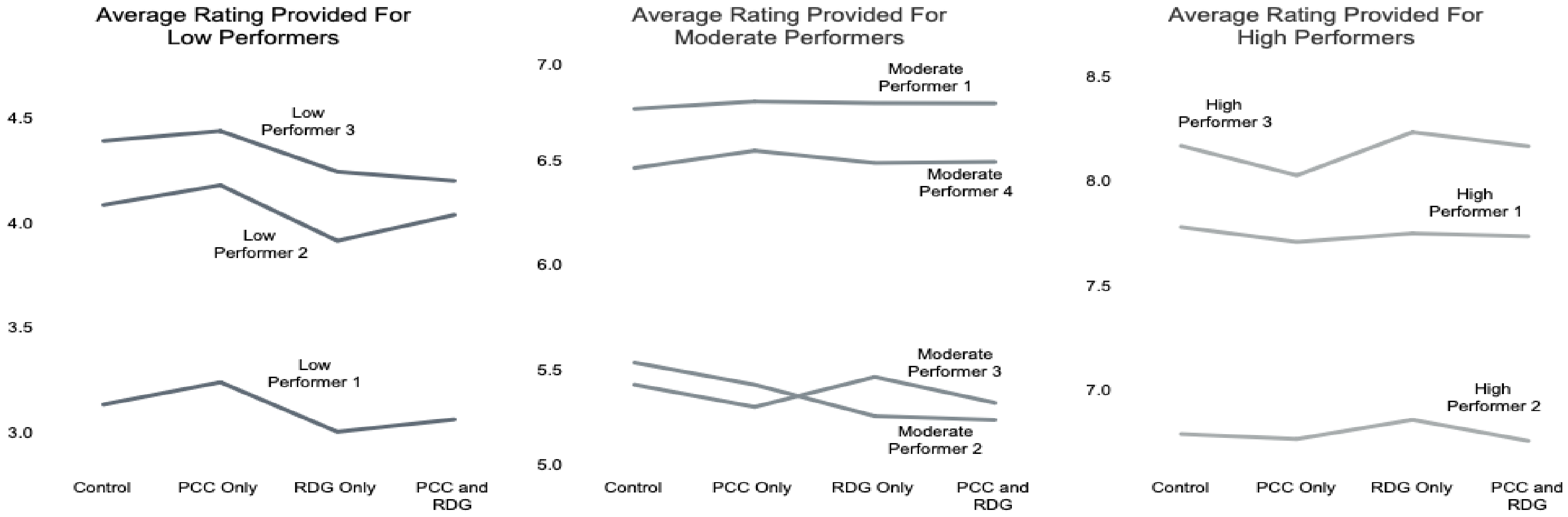
**** p<0.01, ** p<0.05, * p<0.1.*

Panel E– Summary Statistics of the Leniency Bias Rating for Each Performance Level by Condition

						25th		75th	
		N	**Mean**	**Std Dev**	**Minimum**	**Percentile**	**Median**	**Percentile**	**Maximum**
Low Performers									
	Control	85	0.212	0.442	-1.000	0.000	0.000	0.333	1.333
	PCC Only	85	0.294	0.403	-0.333	0.000	0.000	0.667	1.333
	RDG Only	91	0.062	0.415	-1.333	0.000	0.000	0.333	1.000
	PCC and RDG	86	0.109	0.417	-1.667	0.000	0.000	0.333	1.000
Moderate Performers									
	Control	85	-0.200	0.415	-1.250	-0.500	0.000	0.000	0.750
	PCC Only	85	-0.224	0.372	-1.250	-0.500	-0.250	0.000	0.750
	RDG Only	91	-0.242	0.400	-1.500	-0.500	-0.250	0.000	0.750
	PCC and RDG	86	-0.276	0.316	-1.000	-0.500	-0.250	0.000	0.250
High Performers									
	Control	85	0.088	0.429	-0.833	-0.167	0.167	0.500	1.167
	PCC Only	85	0.010	0.332	-0.833	-0.167	-0.167	0.167	0.833
	RDG Only	91	0.123	0.386	-1.167	-0.167	0.167	0.500	1.167
	PCC and RDG	86	0.062	0.422	-0.833	-0.167	0.167	0.167	1.500

Note. In calculating the Leniency Bias measure, the median of the midpoints is taken as the non-biased ratings of that subordinate profile (*Subordinate Midpoint Rating*). Therefore, to evaluate the bias displayed, this *Subordinate Midpoint Rating* is subtracted from the rating given by the participants to create a measure of bias shown by that participant. (See Table 1 - Panel D for the midpoints of each subordinate profile). Statistics are averaged by performance level shown by condition.

Panel F– Graphical Depiction of the Average Rating for Each Profile by Performance Level for Each Condition



Note. Graphs are based on the average of the raw rating scores of each subordinate *Profile*. Graphs are presented by the designed performance level of subordinate *Profiles*.

Table 2 - Correlations

Panel A – Pairwise Correlation of Independent Variables of PCC and RDG with Subordinate Rating by Performance Level and Key Demographic Variables

	PCC (No=0)	RDG (No=0)	Evaluation Experience (No=0)	Gender	Age	Time	Average Leniency	Low Performer	Moderate Performer	High Performer
PCC (No PCC=0)	1.000									
RDG (No RDG=0)	-0.014	1.000								
Evaluation Experience (No=0)	-0.078	0.081	1.000							
Gender (Male=0)	0.031	0.009	-0.103*	1.000						
Age	0.045	0.042	0.229*	-0.045	1.000					
Time	0.112*	0.074	0.097*	-0.009	0.107*	1.000				
Average Leniency	-0.021	-0.093*	-0.090*	-0.006	0.006	-0.008	1.000			
Low Performers	0.078	-0.197*	-0.149*	0.088	-0.008	-0.017	0.713*	1.000		
Moderate Performers	-0.038	-0.062	-0.075	-0.063	0.026	0.004	0.862*	0.451*	1.000	
High Performers	-0.089*	0.056	0.027	-0.029	-0.009	-0.006	0.679*	0.161*	0.436*	1.000

Note. Two-tailed Pearson Correlations of the main independent variables of PCC and RDG with *Evaluation Experience* with key demographic variables and main dependent variables of *Average Leniency* (average leniency of all ten subordinates rated) and the leniency bias displayed on average for *Low Performers, Moderate Performers,* and *High Performers*.
**** p<0.01, ** p<0.05, * p<0.1*

Table 3 - Hypothesis Test

Panel A – Bootstrap Specifications

Sampling Method	Stratified
Number of Samples	1000
Strata Variables	Subordinate #

Panel B - Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	p-value[a]
Intercept	1	3465	1.188	0.276
PCC	1	3465	0.532	0.466
RDG	1	3465	4.890	0.027 **
PCC * RDG	1	3465	0.049	0.825
Evaluation Experience	1	3465	4.785	0.029 **

Note. A bootstrapped restricted maximum likelihood HLM is used to test for the effects of *PCC* and *RDG* on *Average Leniency* with the covariate *Evaluation Experience*. Bootstrap results are based on 1000 stratified bootstrap samples.

[a] Significance is determined based on two-tailed testing.

**** p<0.01, ** p<0.05, * p<0.1*

Table 4 - Hypothesis Testing – Subordinates Partitioned by Performance Level

Panel A – Bootstrap Specifications

Sampling Method	Stratified
Number of Samples	1000
Strata Variables	Subordinate #

Panel B - Type III Tests of Fixed Effects

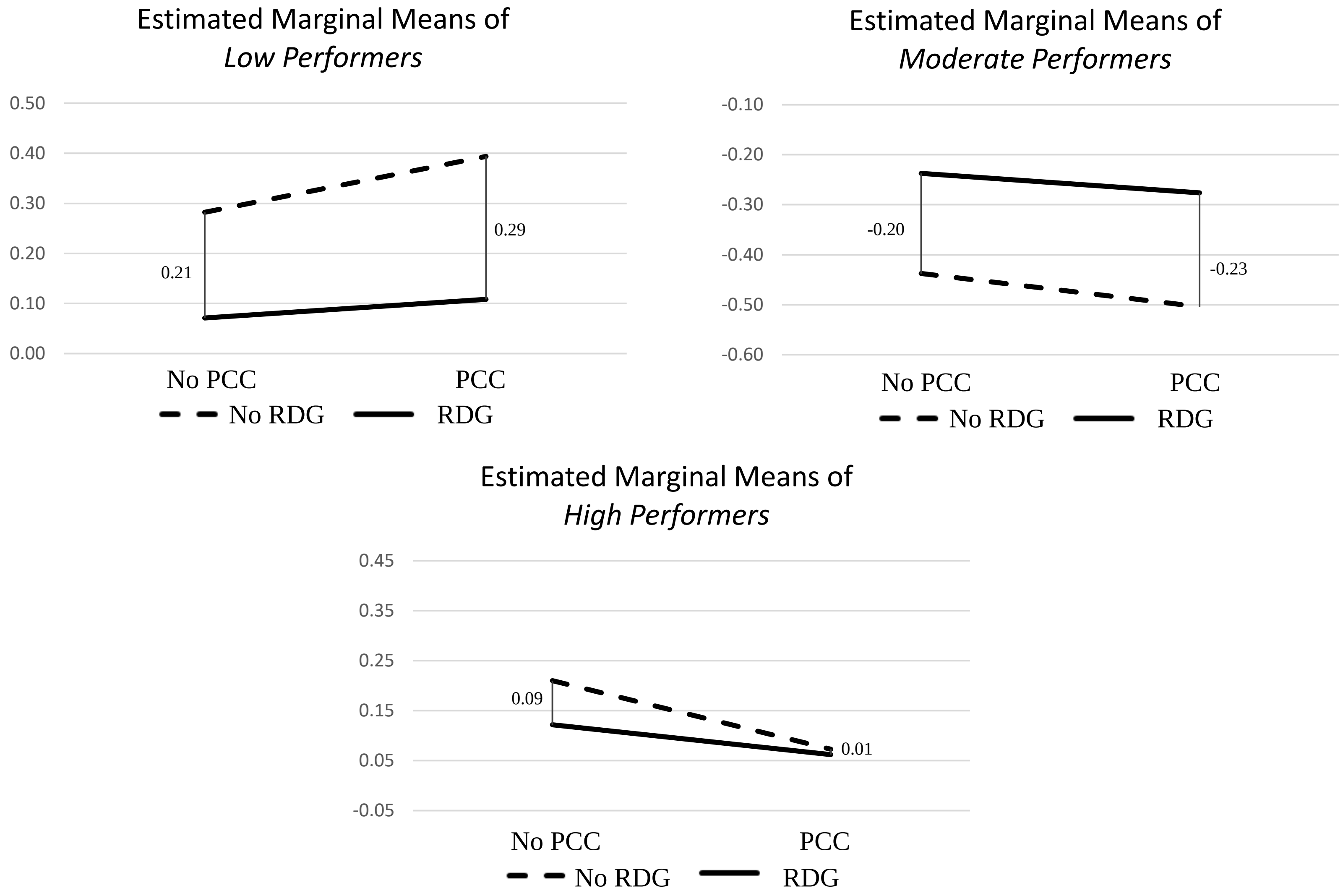
Performance Level		**Numerator**	**Denominator**			
	Source	**df**	**df**	**F**	**p-value**[a]	
Low Performer	Intercept	1	1036	57.204	<0.001	
	PCC	1	1036	2.804	0.094	*
	RDG	1	1036	22.673	<0.001	***
	PCC * RDG	1	1036	0.312	0.576	
	Evaluation Experience	1	1036	10.890	0.001	***
Moderate Performer	Intercept	1	1383	14.883	<0.001	
	PCC	1	1383	0.605	0.437	
	RDG	1	1383	0.998	0.318	
	PCC * RDG	1	1383	0.018	0.894	
	Evaluation Experience	1	1383	1.674	0.196	
High Performer	Intercept	1	1036	2.116	0.146	
	PCC	1	1036	3.496	0.062	*
	RDG	1	1036	1.331	0.249	
	PCC * RDG	1	1036	0.060	0.807	
	Evaluation Experience	1	1036	0.113	0.737	

Note. A bootstrapped restricted maximum likelihood HLM is used to test for the effects of *PCC* and *RDG* on *Low Performers' Leniency Bias*, *Moderate Performers' Leniency Bias* and *High Performers' Leniency Bias*, with the covariate *Evaluation Experience*. Bootstrap results are based on 1000 stratified bootstrap samples. Pairwise comparisons are adjusted for multiple comparisons using Bonferroni with the covariate appearing in the model evaluated at *Evaluation Experience* = 0.79.

[a] Significance is determined based on two-tailed testing.

**** p<0.01, ** p<0.05, * p<0.1*

Panel B – Graphical Depiction of the Average Leniency Bias by Performance Level for Each Condition



Note. Graphs are based on the *Leniency Bias* measure and covariates in the model are evaluated at Evaluation Experience (No=0) at 0.79.

Table 5 - Additional Hypothesis Testing – Planned Contrast Testing – Low Performers

Panel A – ANOVA Leniency Bias in Low Performers' Ratings

			Sum of Squares	df	Mean Square	F	p-value[a]
Between Groups	(Combined)		8.508	3	2.836	9.790	<0.001
	Linear Term	Unweighted	3.772	1	3.772	13.019	<0.001
		Weighted	3.838	1	3.838	13.249	<0.001
		Deviation	4.670	2	2.335	8.060	<0.001
Within Groups			300.409	1037	0.290		
Total			308.916	1040			

Panel B – Contrast Coefficients

Contrast	Control	PCC Only	RDG Only	PCC and RDG
1	1	2	-2	-1
2	1	3	-6	2

Panel C – Planned Contrast Testing

								95% Confidence Interval		
		Contrast	Value of Contrast	Std. Error	t	df	p-value[a]	Lower	Upper	
Leniency Bias Measure	Assumes equal variances	1	0.938	0.232	4.033	1037	<0.001	0.4813	1.3938	***
		2	0.567	0.105	5.394	1037	<0.001	0.3607	0.7732	***
	Does not assume equal variances	1	0.938	0.235	3.997	468.545	<0.001	0.4766	1.3985	***
		2	0.567	0.103	5.491	800.269	<0.001	0.3643	0.7696	***

Note. Planned contrast testing was conducted to analyze the predicted pattern on results for H1, H2, and H3. That is that PCCs will have a main effect of increasing leniency bias, RDG will have a main effect of weakening Leniency Bias, and RDG will weaken the effect of a PCC on leniency bias. For robustness I test two contrasts following the same overall pattern. This model tests the contrasts within the *Low Performer* group. Custom planned contrast coefficients follow predicted patterns of hypothesized results (Buckless & Ravenscroft, 1990; Guggenmos et al., 2018).

[a] Significance is determined based on two-tailed testing.

**** p<0.01, ** p<0.05, * p<0.1*

Table 6 - Additional Hypothesis Testing – Planned Contrast Testing – Moderate Performers

Panel A – ANOVA Leniency Bias in Moderate Performers' Ratings

			Sum of Squares	**df**	**Mean Square**	**F**	**p-value[a]**
Between Groups	(Combined)		1.060	3	0.353	0.560	0.641
	Linear Term	Unweighted	1.044	1	1.044	1.655	0.199
		Weighted	1.041	1	1.041	1.651	0.199
		Deviation	0.019	2	0.009	0.015	0.985
Within Groups			872.902	1384	0.631		
Total			873.962	1387			

Panel B – Contrast Coefficients

Contrast	**Control**	**PCC Only**	**RDG Only**	**PCC and RDG**
1	1	2	-2	-1
2	1	3	-6	2

Panel C – Contrast Tests

								95% Confidence Interval	
		Contrast	**Value of Contrast**	**Std. Error**	**t**	**df**	**p-value[a]**	**Lower**	**Upper**
Leniency Bias	Assumes equal variances	1	0.1126	0.13431	0.839	1384	0.402	-0.1509	0.3761
		2	0.0276	0.29709	0.093	1384	0.926	-0.5552	0.6104
	Does not assume equal variances	1	0.1126	0.13469	0.836	1024.665	0.403	-0.1517	0.3769
		2	0.0276	0.29536	0.094	674.835	0.925	-0.5523	0.6076

Note. Planned contrast testing was conducted to analyze the predicted pattern on results for H1, H2, and H3. That is that PCCs will have a main effect of increasing leniency bias, RDG will have a main effect of weakening Leniency Bias, and RDG will weaken the effect of a PCC on leniency bias. For robustness, I test two contrasts following the same overall pattern. This model tests the contrasts within the *Moderate Performer* group. Custom planned contrast coefficients follow predicted patterns of hypothesized results (Buckless & Ravenscroft, 1990; Guggenmos et al., 2018).

[a] Significance is determined based on two-tailed testing.

**** p<0.01, ** p<0.05, * p<0.1*

Table 7 - Additional Hypothesis Testing – Planned Contrast Testing – High Performers

Panel A – ANOVA of Leniency Bias in High Performers' Ratings

			Sum of		Mean		
			Squares	**df**	**Square**	**F**	**p-value[a]**
Between Groups	(Combined)		1.781	3	0.594	1.707	0.164
	Linear Term	Unweighted	0.015	1	0.015	0.043	0.835
		Weighted	0.018	1	0.018	0.051	0.821
		Deviation	1.763	2	0.882	2.535	0.080
Within Groups			872.902	360.637	1037	0.348	
Total			873.962	362.418	1040		

Panel B – Contrast Coefficients

Contrast	Control	PCC Only	RDG Only	PCC and RDG
1	1	2	-2	-1
2	1	3	-6	2

Panel C – Contrast Tests

								95% Confidence	
			Value of	**Std.**				**Interval**	
		Contrast	**Contrast**	**Error**	**t**	**df**	**p-value[a]**	**Lower**	**Upper**
Leniency Bias	Assumes equal variances	1	-0.1389	0.11612	-1.196	1037	0.232	-0.3668	0.0890
		2	-0.0090	0.25934	-0.035	1037	0.972	-0.5179	0.4999
	Does not assume equal variances	1	-0.1389	0.11671	-1.190	735.357	0.234	-0.3680	0.0902
		2	-0.0090	0.26739	-0.034	427.745	0.973	-0.5346	0.5165

Note. Planned contrast testing was conducted to analyze the predicted pattern on results for H1, H2, and H3. That is that PCCs will have a main effect of increasing leniency bias, RDG will have a main effect of weakening Leniency Bias, and RDG will weaken the effect of a PCC on leniency bias. For robustness I test two contrasts following the same overall pattern. This model tests the contrasts within the *Moderate Performer* group. Custom planned contrast coefficients follow predicted patterns of hypothesized results (Buckless & Ravenscroft, 1990; Guggenmos et al., 2018).

[a] Significance is determined based on two-tailed testing.

****** *$p<0.01$, ** $p<0.05$, * $p<0.1$*

Table 8 – Summary Statistics – Social Comparison, Impression Management and Norms

Panel A - Descriptive Statistics of Social Comparison, Impression Management and Norms by Condition

	Control			PCC Only			RDG Only			PCC and RDG			All Conditions		
	Median	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**	**Median**	**Mean**	**Std Dev**
SC_Think	5	4.071	1.844	4	3.506	1.77	4	4.121	1.825	4	3.756	1.834	4	3.867	1.828
SC_Concern	4	3.541	1.701	4	3.576	1.854	4	3.813	1.653	4	3.965	1.56	4	3.726	1.696
SC_ Interferes	3	3.047	1.640	2	2.918	1.761	3	3.308	1.561	3	3.302	1.681	3	3.147	1.662
IM_DemAbility	4	3.671	1.735	4	4.000	1.753	4	3.846	1.653	4	4.116	1.552	4	3.908	1.675
IM_PeerOpin	3	3.106	1.746	3	3.506	1.688	4	3.582	1.745	4	4.023	1.587	3	3.556	1.717
IM_SubFairPercep	5	5.071	1.437	5	4.800	1.609	5	5.110	1.449	6	5.326	1.443	5	5.078	1.491
IM_SupOpin	4	4.141	1.612	4	3.988	1.842	4	4.033	1.716	4	4.174	1.558	4	4.084	1.68
N_Fairness	6	5.976	1.185	6	6.024	1.165	6	6.099	0.967	6	6.081	1.18	6	6.046	1.122
N_High	3	3.212	1.726	3	2.894	1.705	3	3.209	1.630	3	3.407	1.633	3	3.182	1.676
N_Low	3	3.435	1.546	3	3.306	1.780	4	3.527	1.493	4	3.849	1.598	4	3.530	1.611
N_ConsidPeerOp	4	3.718	1.532	3	3.494	1.702	4	3.835	1.408	4	3.744	1.543	4	3.700	1.546
N_Dist	4	3.553	1.500	4	3.435	1.756	4	3.769	1.578	3	3.291	1.379	4	3.516	1.562
N_Accuracy	7	6.4	1.157	7	6.482	0.854	7	6.352	1.037	7	6.605	0.724	7	6.458	0.959
PCC_AdjDown	.	.	.	4	3.494	1.702	.	.	.	4	3.651	1.54	4	3.573	1.619
PCC_Negotiate	.	.	.	4	3.894	1.626	.	.	.	4	4.07	1.525	4	3.982	1.574
N			**85**			**85**			**91**			**86**			**347**

Note. Summary descriptive statistics for all PEQ questions aiming to capture measures of the theoretical constructs. See Appendix E – Panel A for all PEQ questions. Statistics are shown by condition and as the total for all conditions.

Table 9 – Factor Principal Components Analysis

Panel A – Total Variance Explained

	Rotation Sums of Squared Loadings		
Component	**Total**	**% of Variance**	**Cumulative %**
1	2.449	20.410	20.410
2	2.424	20.204	40.614
3	1.595	13.289	53.903
4	1.435	11.957	65.860

Panel B - Rotated Component Matrix

	Component			
	1	**2**	**3**	**4**
N_High	**0.830**			
N_Low	**0.823**			
N_Dist	**0.708**			
N_Peer	**0.569**	0.544		
IM_DemAbil		**0.855**		
IM_PeersOpin		**0.757**	0.330	
IM_Impress		**0.663**		
PCC_Down			**0.819**	
PCC_Negotiation			**0.814**	
N_Fair				**0.801**
N_Acc	-0.318			**0.653**
IM_Fair		0.487		**0.543**

Note. Components were extracted using a Principle Component Analysis with a Varimax Rotation with Kaiser Normalization. Rotation converged in 5 iterations. Scores below 0.300 were suppressed for reporting purposes.

****** p<0.01, ** p<0.05, * p<0.1*

Table 11 – Testing of Theoretical Constructs

Panel A – ANCOVA Test of Between-Subjects Effects on Impression Management (IM)

Source	Type III Sum of Squares	df	Mean Square	F	Sig. (two-tailed)	
Corrected Model	128.053	5	25.611	16.452	<0.001	
Intercept	97.360	1	97.360	62.543	<0.001	
PCC	7.599	1	7.599	4.881	0.028	**
RDG	1.261	1	1.261	0.810	0.369	
PCC * RDG	3.654E-5	1	3.654E-5	0.000	0.996	
Dist_Norm	118.568	1	118.568	76.166	<0.001	***
Evaluation Experience	0.747	1	0.747	0.480	0.489	
Error	530.832	341	1.839			
Total	5800.111	347				
Corrected Total	658.885	346				

Note. ANCOVA testing of PCC and RDG on the theoretical construct of Impression Management and including testing for the hypothesized relationship in H3 of the distribution accuracy norm on Impression Management Behaviours.

R Squared = 0.194 (Adjusted R Squared = 0.183)

***p<0.01, **p<0.05, *p<0.1

Panel B – ANCOVA Test of Between-Subjects Effects on Distribution Norms (Dist_Norm)

Source	Type III Sum of Squares	df	Mean Square	F	Sig. (two-tailed)	
Corrected Model	9.307	4	2.327	1.652	0.161	
Intercept	976.310	1	976.310	693.068	<0.001	
Evaluation Experience	4.243	1	4.243	3.012	0.084	*
PCC	1.282	1	1.282	0.910	0.341	
RDG	4.014	1	4.014	2.850	0.092	*
PCC * RDG	0.722	1	0.722	0.512	0.475	
Error	481.768	342	1.409			
Total	4698.188	347				
Corrected Total	491.075	346				

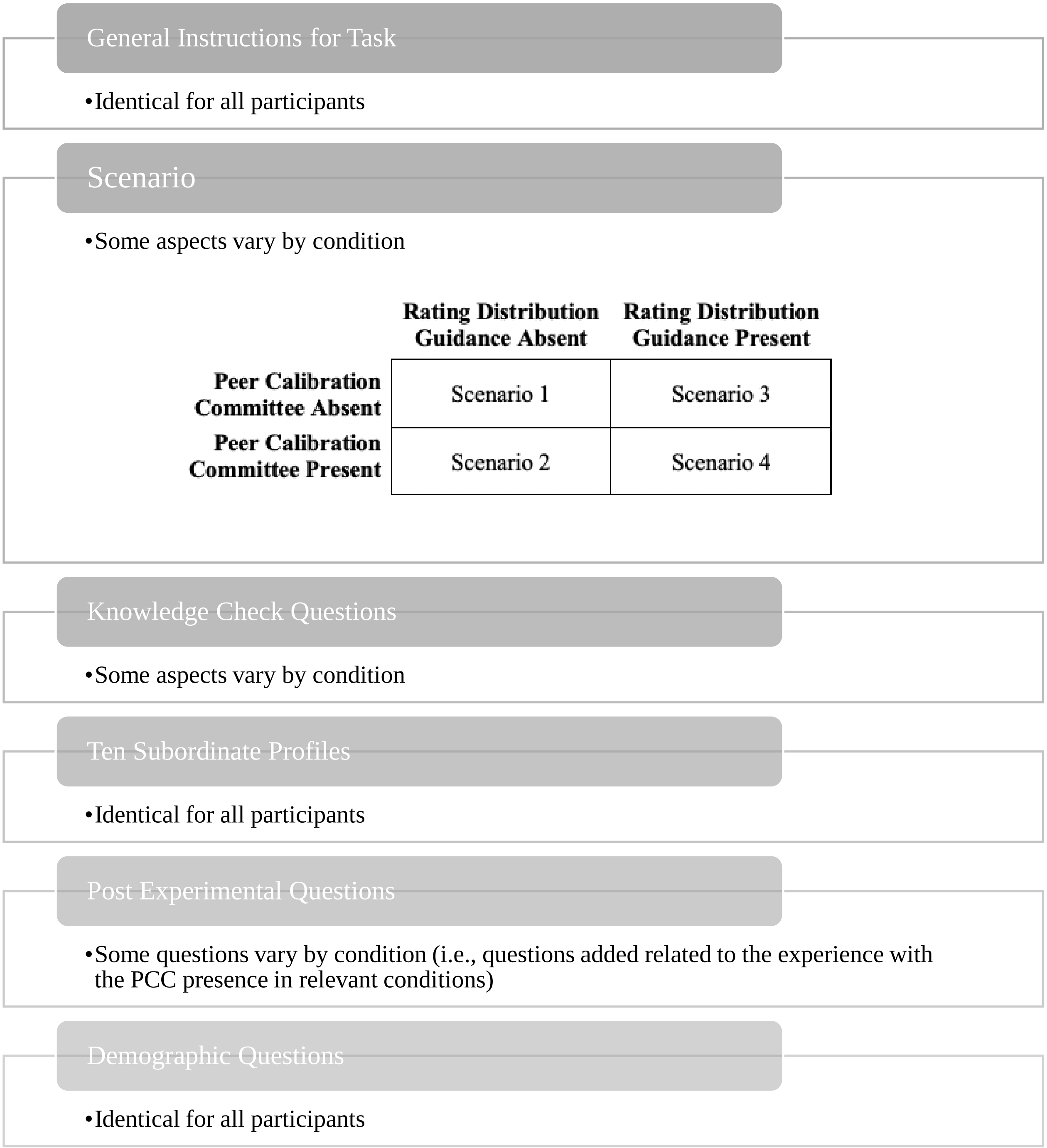
Note. ANCOVA testing of PCC and RDG on the theoretical construct of Distribution Accuracy Norm.

R Squared = 0.019 (Adjusted R Squared = 0.007)

***p<0.01, **p<0.05, *p<0.1

APPENDICES

Appendix A - Experimental Design Overview



General Instructions for Task

- Identical for all participants

Scenario

- Some aspects vary by condition

	Rating Distribution Guidance Absent	Rating Distribution Guidance Present
Peer Calibration Committee Absent	Scenario 1	Scenario 3
Peer Calibration Committee Present	Scenario 2	Scenario 4

Knowledge Check Questions

- Some aspects vary by condition

Ten Subordinate Profiles

- Identical for all participants

Post Experimental Questions

- Some questions vary by condition (i.e., questions added related to the experience with the PCC presence in relevant conditions)

Demographic Questions

- Identical for all participants

Appendix B - General Instructions [50,51]

For all experimental conditions

Thank you for participating in this study. The purpose of the study is to investigate how managers make judgments and decisions. There are no right or wrong answers to the questions you will be asked. Your participation should take approximately 15 minutes.

For the purposes of this study, you are asked to assume the role of a team supervisor at Peninsula Industries, Inc. (Peninsula). Your task is to: (1) review information related to your team's performance; and (2) determine each employee's performance rating for the year.

The case details you will read are partial and are not intended to be entirely representative of all the information you would normally have if you were "on the job." Furthermore, the decisions you will be making do not necessarily include all of the decisions you would normally make. The objective of this study is to convey to you enough information so that you are able to make the requested decisions. For purposes of this study, please base your judgments only on the information provided.

Thank you again for your participation.

[50] Adapted from (Bailey et al., 2011)

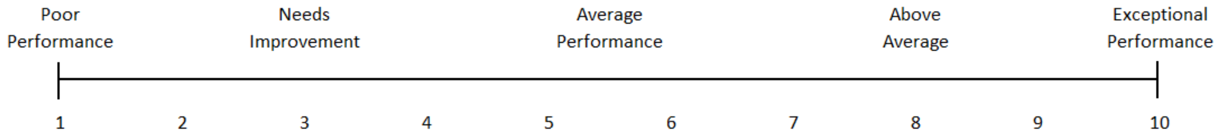
[51] To use language more familiar to participants, the word "employee" is used throughout the instrument rather than subordinate.

Appendix C – Scenarios [52,53]

Scenario 4 – Presence of PCC, Presence of RDG

Your Role

Assume you are a team supervisor at Peninsula Industries, Inc. Your firm is currently engaging in its annual performance review process, and **you have been asked to evaluate your team of 10 employees**. As part of the annual performance review process, you are to rate each of your employees on a scale from 1-10. The following rating scale has been provided to you for assessing your employees:



[Page Break]

The firm believes all supervisors should provide accurate ratings as doing so provides important feedback to employees about their performance.

[PCC PRESENT CONDITIONS ONLY]

*[Therefore, assume after you complete the employee ratings, the next step will be to meet with a group of **4** other supervisors in your department. This committee will review the ratings of each employee and calibrate ratings across supervisors. **The goal of this committee is to reduce any differences in employee ratings across the supervisors.**]*

[Page Break]

{RDG PRESENT CONDITIONS ONLY}

{Human Resources has provided guidance approved by the CEO that, on average, between all supervisors and departments, 20% of employees should be rated 8-10, 20% of employees should be 1-3, with the remaining 60% being rated between 4-7.}

[Page Break]

Assume after you complete the evaluations, you will be responsible for discussing them with your employees. **You would then submit their evaluations to HR for inclusion in each employee's file**. A summary of your team's ratings would also be sent to your supervisor as a part of his/her resources for assessing your own performance.

[52] Text that differs between conditions is italicized and red to identify adaptation.

[53] Knowledge check questions are examples based on this condition. Knowledge check questions vary between conditions as appropriate.

The firm believes how an employee performs reflects both the employee's ability and effort and the supervisor's ability to bring out the best in their employees.

[Page Break]

[PCC PRESENT CONDITIONS ONLY]

Assume the following about the calibration committee members:

- *They all manage teams in your department*
- *They all have similar work experience to yours*
- *They all have employee teams that are similar to yours*
- *They all present their employee ratings to the committee for review*
- *You frequently interact and work with each of them*
- *You care a great deal that they think you are a good supervisor*

[Page Break]

Test your knowledge about your job:

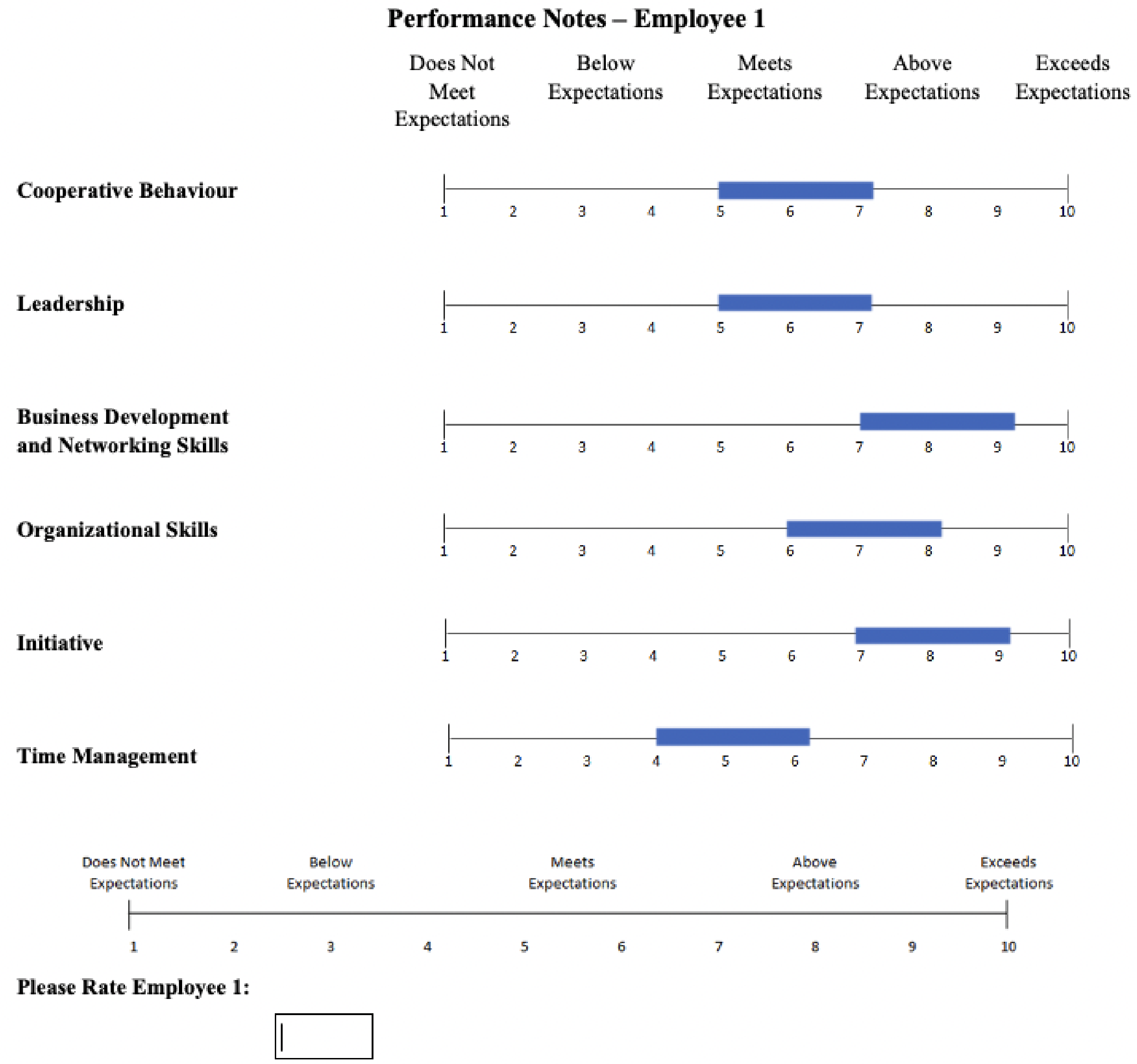
1. *True or False: You will be responsible for assessing your employee's performance.*
2. *True or False: The rating guidance provided is that the average of all employees in the firm should be:*

   20% of employees should be 1-3,
   60% being rated between 4-7, and
   20% of employees should be rated 8-10.
3. *True or False: You will be submitting your ratings to the committee.*
4. *True or False: The members of the committee will be reviewing your ratings.*
5. *True or False: The opinions of the supervisors on the committee matter to you.*

[Page Break]

Appendix D - Subordinate Profiles

Panel A – Example of the Subordinate Profile for Employee 1



Note. Participants are instructed that the blue bars represent their preliminary assessment of the subordinate's performance in each category.

Panel B – Instructions and Full Overview of all Ten Subordinate Profiles

On the following pages, you will find an overview of the performance information that you have collected on six different categories for each of the ten employees on your team. The firm identified these categories as the key criteria for an employee's performance and believes all six categories should be equally weighted.

Please assume that the blue bar present on the rating scale for each category indicates your own preliminary assessment of that employee's performance for that category.

Please provide an overall rating for each employee based on each profile presented.

– – – – – – – – – – – – – [Page Break] - – – – – – – – – – – – – -

Performance Notes – Employee 1

Does Not Meet Expectations

Below Expectations

Meets Expectations

Above Expectations

Exceed Expectati

Cooperative Behaviour

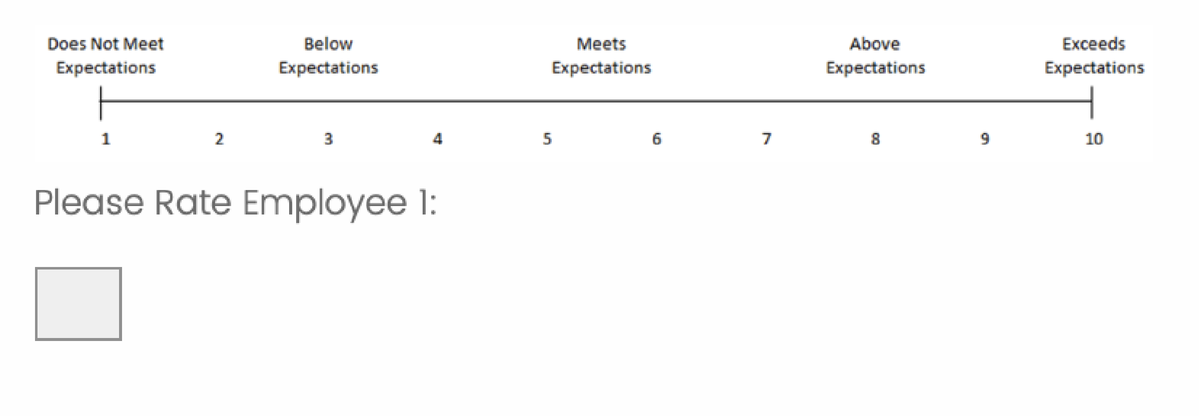
Leadership

Business Development and Networking Skills

Organizational Skills

Initiative

Time Management



Please Rate Employee 1:

□

– – – – – – – – – – – [Page Break] – – – – – – – – – – –

Performance Notes – Employee 2

Does Not Meet Expectations | Below Expectations | Meets Expectations | Above Expectations | Exceeds Expectations

Cooperative Behaviour

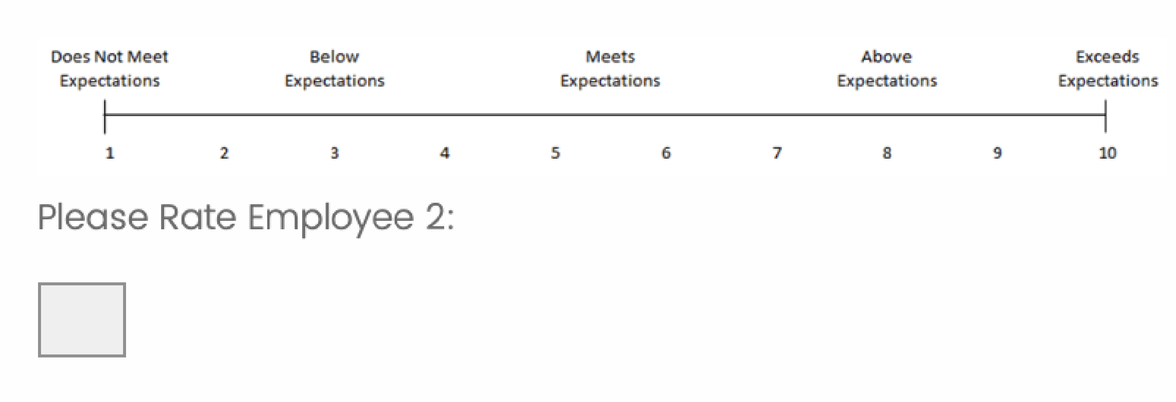
Leadership

Business Development and Networking Skills

Organizational Skills

Initiative

Time Management



Does Not Meet Expectations | Below Expectations | Meets Expectations | Above Expectations | Exceeds Expectations

1 2 3 4 5 6 7 8 9 10

Please Rate Employee 2:

□

[Page Break]

Performance Notes – Employee 3

Does Not Meet Expectations | Below Expectations | Meets Expectations | Above Expectations | Exceeds Expectations

Cooperative Behaviour

1
2
3
4
5
6
7
8
9
10

Leadership

1
2
3
4
5
6
7
8
9
10

Business Development and Networking Skills

1
2
3
4
5
6
7
8
9
10

Organizational Skills

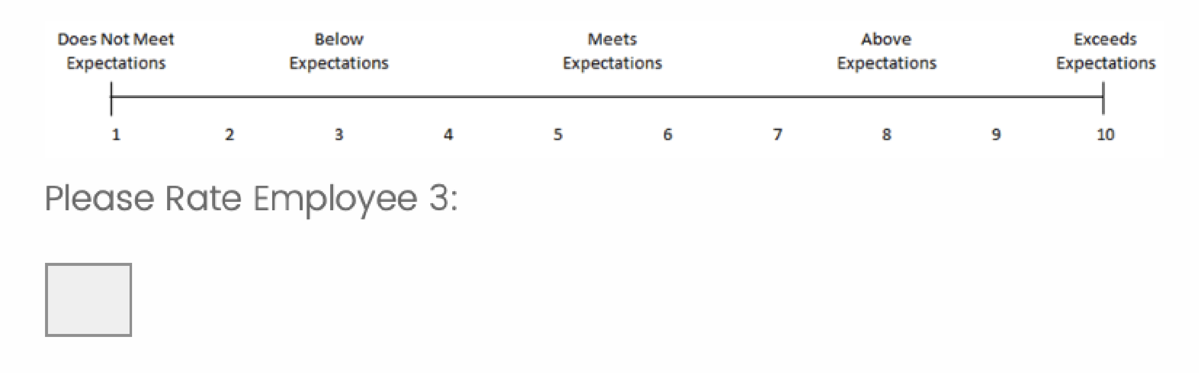
1
2
3
4
5
6
7
8
9
10

Initiative

1
2
3
4
5
6
7
8
9
10

Time Management

1
2
3
4
5
6
7
8
9
10



Please Rate Employee 3:

□

[Page Break]

Performance Notes – Employee 4

Does Not Meet Expectations	Below Expectations	Meets Expectations	Above Expectations	Exceeds Expectations

Cooperative Behaviour

1
2
3
4
5
6
7
8
9
10

Leadership

1
2
3
4
5
6
7
8
9
10

Business Development and Networking Skills

1
2
3
4
5
6
7
8
9
10

Organizational Skills

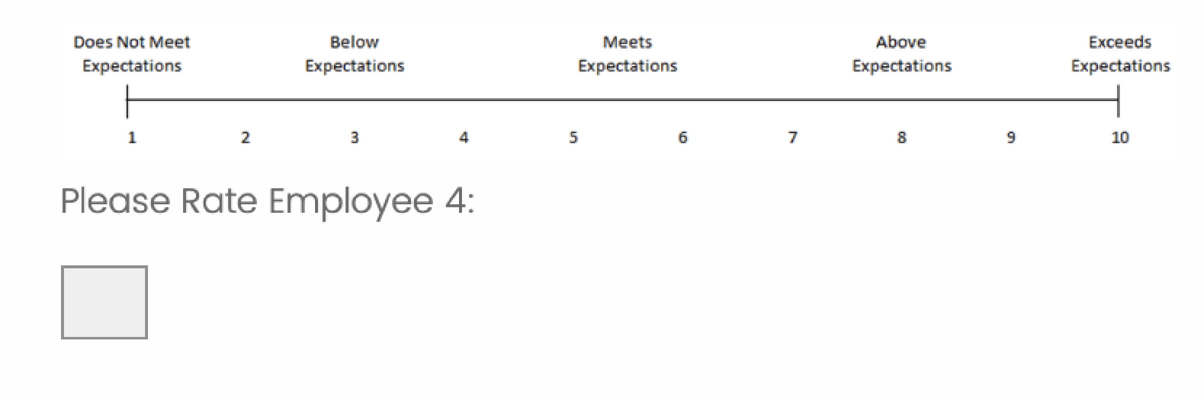
1
2
3
4
5
6
7
8
9
10

Initiative

1
2
3
4
5
6
7
8
9
10

Time Management

1
2
3
4
5
6
7
8
9
10



Please Rate Employee 4:

[Page Break]

Performance Notes – Employee 5

Does Not Meet Expectations	Below Expectations	Meets Expectations	Above Expectations	Exceeds Expectations

Cooperative Behaviour

1 2 3 4 5 6 7 8 9 10

Leadership

1 2 3 4 5 6 7 8 9 10

Business Development and Networking Skills

1 2 3 4 5 6 7 8 9 10

Organizational Skills

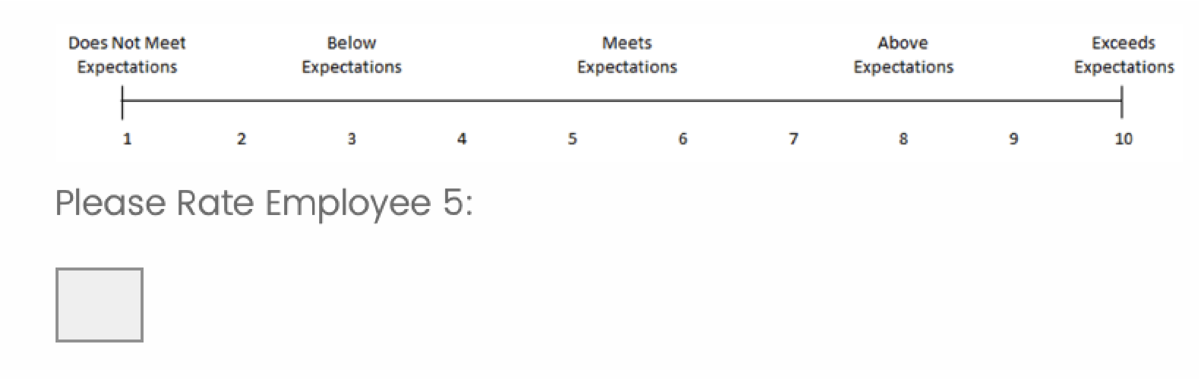
1 2 3 4 5 6 7 8 9 10

Initiative

1 2 3 4 5 6 7 8 9 10

Time Management

1 2 3 4 5 6 7 8 9 10



Does Not Meet Expectations — Below Expectations — Meets Expectations — Above Expectations — Exceeds Expectations

1 2 3 4 5 6 7 8 9 10

Please Rate Employee 5:

□

[Page Break]

Performance Notes – Employee 6

Does Not Meet Expectations	Below Expectations	Meets Expectations	Above Expectations	Exceeds Expectations

Cooperative Behaviour

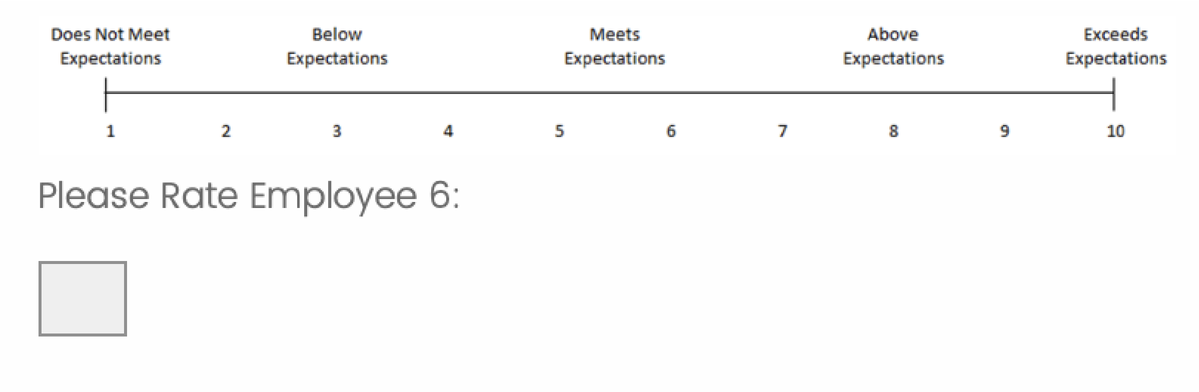
Leadership

Business Development and Networking Skills

Organizational Skills

Initiative

Time Management



Please Rate Employee 6:

□

[Page Break]

Performance Notes – Employee 7

Does Not Meet Expectations | Below Expectations | Meets Expectations | Above Expectations | Exceeds Expectations

Cooperative Behaviour

1
2
3
4
5
6
7
8
9
10

Leadership

1
2
3
4
5
6
7
8
9
10

Business Development and Networking Skills

1
2
3
4
5
6
7
8
9
10

Organizational Skills

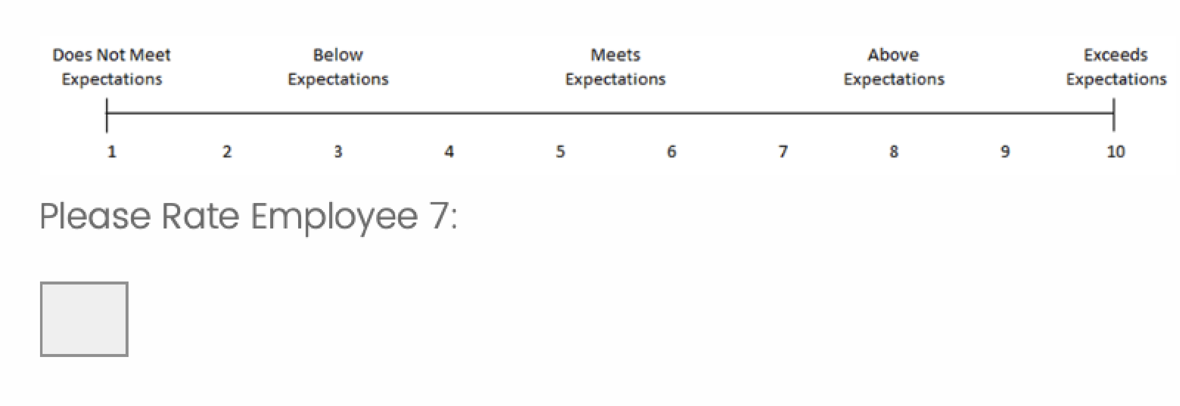
1
2
3
4
5
6
7
8
9
10

Initiative

1
2
3
4
5
6
7
8
9
10

Time Management

1
2
3
4
5
6
7
8
9
10



Please Rate Employee 7:

□

[Page Break]

Performance Notes – Employee 8

Does Not Meet Expectations	Below Expectations	Meets Expectations	Above Expectations	Exceeds Expectations

Cooperative Behaviour

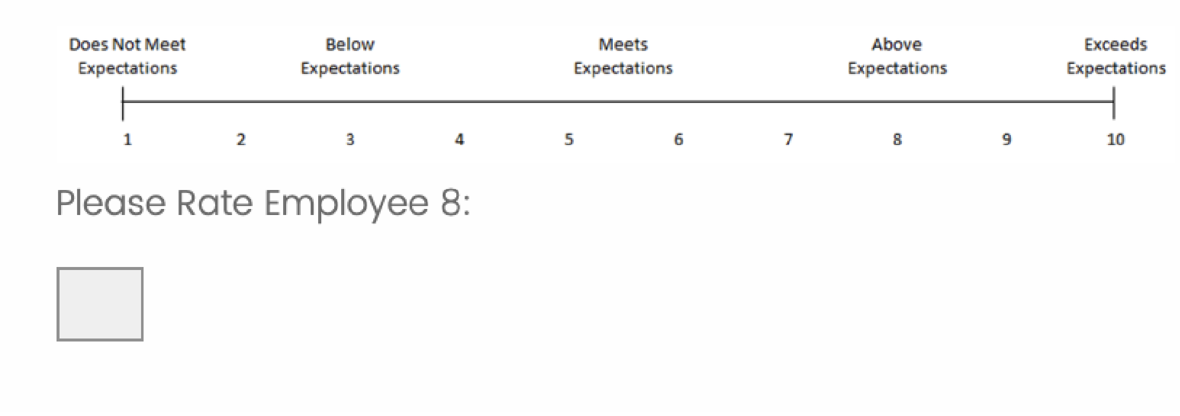
Leadership

Business Development and Networking Skills

Organizational Skills

Initiative

Time Management



Please Rate Employee 8:

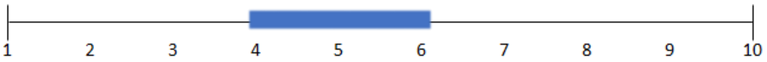
□

[Page Break]

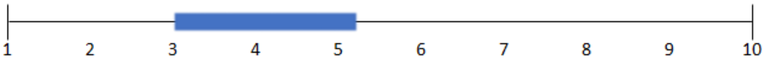
Performance Notes – Employee 9

Does Not Meet Expectations	Below Expectations	Meets Expectations	Above Expectations	Exceeds Expectations

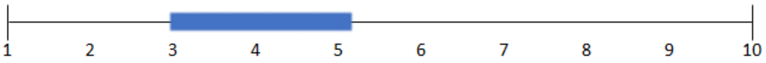
Cooperative Behaviour



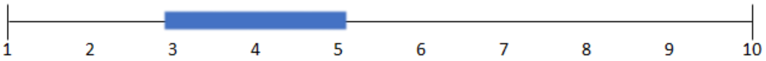
Leadership



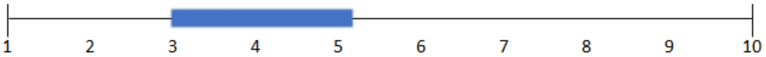
Business Development and Networking Skills



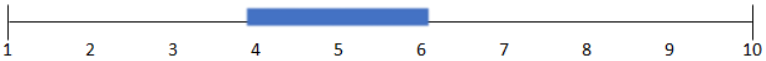
Organizational Skills

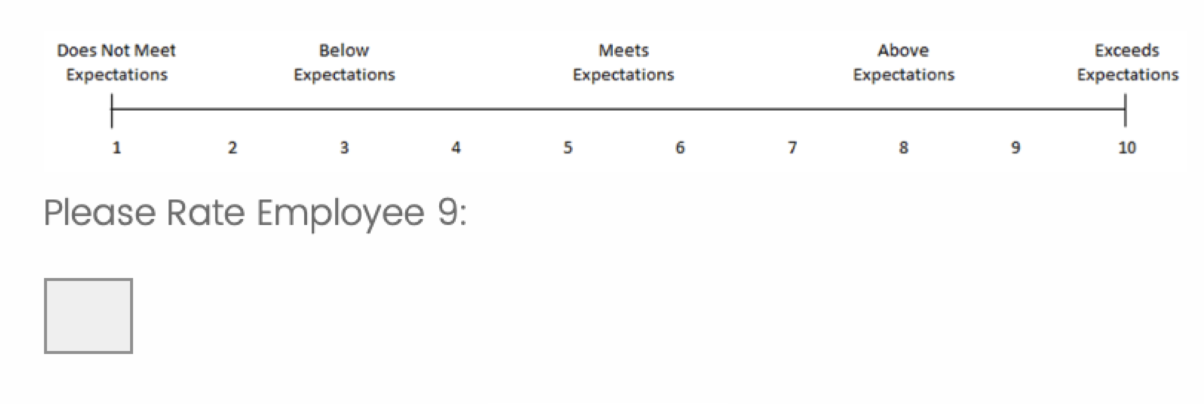


Initiative



Time Management





Please Rate Employee 9:

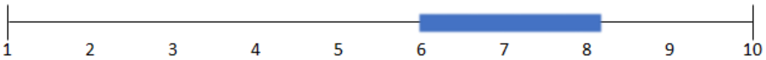
□

[Page Break]

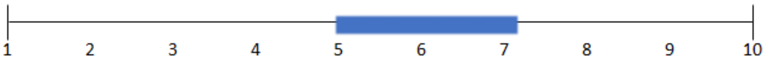
Performance Notes – Employee 10

Does Not Meet Expectations	Below Expectations	Meets Expectations	Above Expectations	Exceeds Expectations

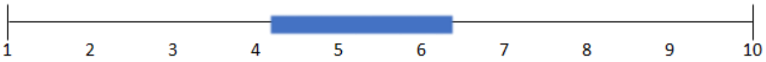
Cooperative Behaviour



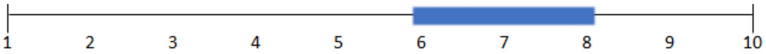
Leadership



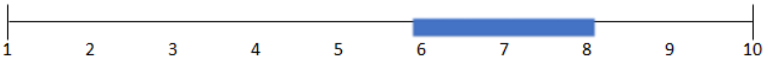
Business Development and Networking Skills



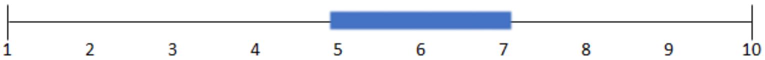
Organizational Skills

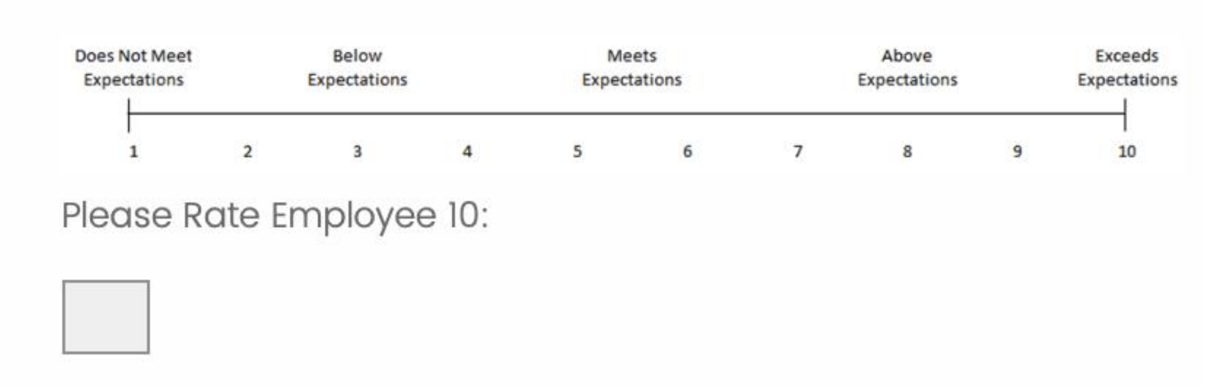


Initiative



Time Management





Please Rate Employee 10:

□

[Page Break]

Appendix E – Post-Experimental and Demographic Questions

Panel A – Post-Experimental Questions

Each question appears on a separate page, and participants provide their responses on the following scale:

○	○	○	○	○	○	○
1	2	3	4	5	6	7
Not at all						To a great extent

[Instructions] Please indicate how much you agree with each statement below using the provided scale. There are no right or wrong answers.

Q1 [SC_Think] To what extent did you think about your employees' ratings relative to other supervisors' employee ratings when completing your evaluations?

Q2 [SC_Concern] To what extent were you concerned about how your employee ratings compared to other supervisors' employee ratings when you rated your employees?

Q3 [SC_ Interfere] To what extent did thinking about comparing your ratings with other supervisors' ratings interfere with your ability to concentrate when you rated your employees?

Q4 [IM – Task is the target for IM] To what extent did you consider your employee ratings as being an opportunity to demonstrate your ability as a supervisor to your peers when you rated your employees?

Q5 [IM – Used task for IM] To what extent did you consider other supervisors' opinions about your abilities as a supervisor when you rated your employees?

Q6 [IM – Other Parties] To what extent did you consider how fair your employees would think their ratings were when you rated your employees?

Q7 [Attention Check] To what extent did you consider how important it was to pay attention to the questions in the study? Please select two for the answer to this question.

Q8 [IM – Other Party] To what extent did you consider your own supervisor's opinion of your ability as a supervisor when you rated your employees?

[Questions 9 and 10 only appear in Conditions where a PCC is present]

Q9 [PCC_AdjDown] To what extent did you consider how the calibration committee might adjust your ratings downwards when rating your employees?

Q10 [PCC_Negotiate] To what extent did you consider whether you might have to negotiate your ratings in the calibration committee when rating your employees?

Q11 [Confirm_IM] (to ask about the direction of bias/ask about IM in another way) To what extent would other supervisors' opinions of your abilities as a supervisor increase the ratings you give your employees?

Q12 [Norm_Fairness] To what extent do you believe a manager should consider the fairness of the ratings when evaluating employees?

Q13 [Norm_High] To what extent do you believe a manager should NOT provide too many high ratings when evaluating a group of employees?

Q14 [Norm_Low] To what extent do you believe a manager should NOT provide too many low ratings when evaluating a group of employees?

Q15 [Norm_ConsidPeerOp] – Consider Peers' Opinions] To what extent do you believe a manager should consider how their peers will perceive the ratings they provide?

Q16 [Norm_WideDist] To what extent do you believe a manager should provide a wide distribution of ratings when evaluating employees?

Q17 [Norm_Accuracy] To what extent do you believe a manager should provide the most accurate ratings possible for employees?

Panel B - Demographic Questions

Q1 Gender (please check): Male ____ Female ____ Non-binary ____ Another Gender Identity ____ Prefer not to say ____

Q2 What is your age? ______

Q3 Do you have any professional certifications? Yes____ No____

[If the answer is "Yes" then the following question is displayed]
Q3b Please describe which professional certifications you have (CPA, CFE, etc.) ______

Q4 Have you ever participated in a committee of managers whose goal was to discuss and calibrate performance evaluation ratings? Yes____ No____

Q5 Your most recent work experience is most closely related to which of the following areas (check all that apply):
____ Accounting, Auditing, or Taxation
____ Information systems/technology
____ Finance, banking, or investing
____ Marketing, or sales
____ General management, or personnel
____ Engineering
____ Medical

Q6 How many employees have you supervised at one time? ______

Q7 Have you ever evaluated the performance of employees who reported to you? Yes____ No____

Appendix F - Pilot Test One

Panel A – Pilot Test One Discussion

Overview

To examine whether the text used in my PCC scenario can induce social comparison in the online participant pool, I perform a 3 x 1 between-subjects experiment. I test three conditions for my scenario (Panel B): 1) PCC absent [*No_PCC*], 2) PCC present with high social comparison factors [*PCC_SC_HIGH],* 3) PCC present with low social comparison factors [*PCC_SC_Low*].

Pilot Testing Social Comparison

Prior research finds successful inducement of social comparison in laboratory-based experiments (e.g., Hannan et al., 2013; Tafkov, 2013). To my knowledge, this has not yet been examined using an online vignette experiment. There are benefits to online participants, such as access to a broader pool of participants and participants of a specific skill set (A. Farrell et al., 2017; M. Farrell & Sweeney, 2021; Bentley, 2019; Buchheit et al., 2018, 2019). For example, by using an online subject pool, I am able to access participants with supervisory experience, which brings beneficial experience and external validity to my study (see Section 4.3 for further discussion of participants used in my main experiment). However, the techniques used in prior studies to enhance social comparison require participants to be present in the same room (e.g., Hannan et al., 2013; Tafkov, 2013). Online experiments remove the possibility of having participants complete the task simultaneously in the same room, and, thereby, create a greater sense of completing the task in isolation. This may limit the feeling of social comparability. Therefore, I conducted a pilot test to assess whether the information provided in my PCC present condition effectively invokes a social comparison response in an online labour pool.

Task Details

I design my task to contain information invoking the three key factors that strengthen social comparison and its relation to behaviour—task similarity, differences in task performance result from controllable factors, and the comparison group is important to the individual. In all conditions, RDG is absent, as testing RDG is not a goal of this pilot test. My *NoPCC* condition serves as a control condition for this pilot test. In this condition, I present the same details I do for my final experiment when both PCC and RDG are absent (Panel B). The *PCC_SC_HIGH* condition includes the PCC details I intend to use in my final experiment design (Panel B). Each item is included to strengthen one or more of the factors related to social comparison. Conversely, in the PCC_SC_Low condition, I provide directly opposing information about the PCC to test whether PCC composition can also reduce social comparison from the baseline control condition. By including a condition for both the enhancement and reduction of social comparison, I can more thoroughly examine social comparison in an online environment.

Panel B of this appendix provides the scenario used in Pilot Test One; including the specific details provided for the *PCC_SC_HIGH* and *PCC_SC_LOW*. The statements regarding the PCC (using *PCC_SC_HIGH* for illustration), align with the social comparison factors as follows:

Scenario Design - PCC Committee Details and Social Comparison

Social Comparison Factor	PCC Committee Information Provided	Design Intention
Task Similarity	*“The other committee members all manage teams in your department”* “*All the teams of the supervisors in the committee are similar to your own”*	By including these details I seek to communicate that all the teams and supervisors are all working toward common goals and on a common business area, increasing the feeling of a general role and task similarity with other PCC supervisors.
	“All supervisors present their ratings to the committee for review”	By including this information, I communicate that the specific task for all PCC members is the same: 1) that all PCC members are all rating their subordinates, and 2) that all PCC members are all presenting those ratings for discussion at the PCC meeting.
Differences in task performance result from controllable factors	*“The other committee members all manage teams in your department”* *“All the teams of the supervisors in the committee are similar to your own”* *“These supervisors have similar experience to your own”*	By highlighting that the department, the teams, and the supervisors’ experience levels are all similar, I seek to isolate that: 1) the rating task itself is the key task being evaluated and, 2) that it is not potentially uncontrollable factors (such as department performance or team size) causing differences in subordinate performance ratings.
	From main scenario components: *“[…]the firm believes how an employee performs reflects both the employee’s ability and effort and the supervisor’s ability to bring out the best in their employees.”.*	I further emphasize that the differences in task performance (i.e., subordinate ratings) are from controllable factors using the information from the base scenario (information provided in all conditions). For example, that “[…the firm believes how an employee performs reflects both the employee’s ability and effort and the supervisor’s ability to bring out the best in their employees.]”.
The comparison group is important to the individual	*“The other committee members all manage teams in your department”* *“You frequently interact and work with the other supervisors on the committee”* *“You care if the other committee members think you are a good supervisor”*	These points are all intended to illustrate that the other committee members are individuals that the participant interacts with frequently and truly are their ‘peers’ in the organization. This is designed to communicate the PCC member’s importance to the participant within the scenario.

Participants

As the purpose is to test social comparison in an online setting, I conduct this pilot study with a group of Amazon Mechanical Turk (MTurk) workers as participants. I pre-screen participants for a minimum of an undergraduate (associate or bachelor's) degree, experience supervising subordinates (to ensure that participants have the necessary work experience and that they are similar to those I wish to use in my final experiment), and a 95% HIT approval rating (to increase the likelihood that participants complete the task effortfully).[54,55] Two hundred and twenty-five MTurk workers participate in this pilot study. I remove 18 participants from the analysis due to their poor-quality responses.[56]

Process Measures – Dependent Variable

After participants read the scenarios, they then respond to a series of questions which primarily measure participants' social comparison and impression management concerns. I base these questions on those I use in my main experiment; however, I adapt these questions to indicate that the participant should imagine they will rate a team of subordinates as they do not actually conduct the ratings in this pilot (see Appendix E – Panel A – Questions 1-8). For example, in my main study, I ask, *"To what extent did thinking about comparing your ratings with other supervisors' ratings interfere with your ability to concentrate when you rated your employee?*". However, in the pilot study, I ask, *"To what extent will thinking about comparing your ratings with other supervisors' ratings interfere with your ability to concentrate on your evaluation?"*.

To develop my primary dependent variable I average responses to five questions from my post-experimental questionnaire (Appendix E – Panel A – Questions 1-5) to form a single

[54] I use a different online labour pool for my pilot (MTurk) than I do for my main experiment (Prolific) to increase my access to potential participants. Peer et al.(2017) tests the use of Prolific, MTurk and student data, finding all three participant pools successfully replicate a set of prior findings. Therefore, even though I use a different online labour pool for my pilot than my pilot, I expect similar results between the two.

[55] A HIT is a worker task on the MTurk platform (i.e., this experiment would be a HIT on MTurk). A 95% approval rating indicates that 95% of the time, the MTurk worker has had their submitted work accepted by the HIT provider. A HIT provider may choose to reject the submission of an MTurk worker if they do not follow the instructions of the HIT or do not provide quality work. A rejection will affect the MTurk worker's HIT approval rating. The HIT approval rating is a quality control measure on MTurk.

[56] Participant responses were assessed on three key metrics to determine data quality. Participant removal occurred because (1) Qualtrics identified them as duplicate respondents, (2) they provided identical responses across all scaled questions (11 participants), or (3) it took them significantly less time to complete the study than it did other participants (one participant took less than one minute to complete the study whereas the average participant took over five minutes).

composite measure of social comparison (SCIM).[57] These questions are developed based on prior research (Tafkov, 2013; Webb et al., 2010) to measure social comparison and impression management (see additional discussion in section 4.6.2 of my document).

Overall, I expect that the *PCC_SC_HIGH* will have a higher *SCIM* score than in both the *No_PCC* and the *PCC_SC_LOW* conditions; indicating participants felt the most social comparison pressure in the *PCC_SC_HIGH* condition. I do not have specific expectations regarding the differences between the *PCC_SC_LOW* and the *No_PCC* condition.

Pilot Test One Results and Conclusion

Prior to conducting my analysis of how each condition affects the social comparison felt by participants, I first conduct a component factor analysis on my SCIM variables. This analysis confirms my five questions SCIM questions reduce to a single dimension (untabulated; eigenvalue of 3.00, explaining 59.99% of the variance).

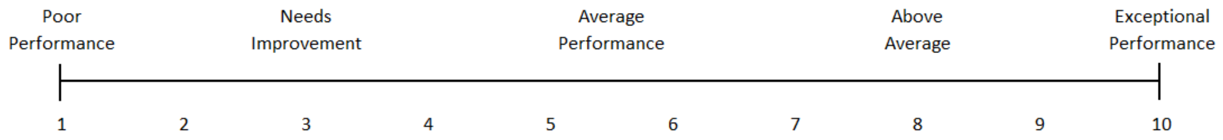
Results confirm my expectations (Panel B and C), showing that social comparison in the *PCC_SC_High* condition ($\mu = 4.68$) is significantly greater than in both the *NoPCC* ($\mu = 3.95$, $p = 0.002$; two-tailed) and the *PCC_SC_Low* conditions ($\mu = 3.75$, $p < 0.001$; two-tailed). Furthermore, I find that *SCIM* was not significantly different between the *NoPCC* ($\mu = 3.95$) and PCC_SC_LOW ($\mu = 3.75$) conditions ($p = 0.399$; two-tailed). Thus, the *PCC_SC_HIGH* setting induced social comparison, and this analysis provides evidence that my design and online participants are suitable for testing my theory.

[57] Questions 6 and 8 from my post-experimental questionnaire ask about how much they believe parties other than the PCC (i.e., the supervisor's subordinates and the supervisor's superior respectively) might factor into their evaluations. Question 7 is an attention check question asking for a specified response of "2".

Panel B - Scenario for Pilot Test One

Your Job

Assume you are a team supervisor at Peninsula Industries, Inc. Your firm is currently engaging in its annual review process, and you have been asked to evaluate your team of 10 employees. As part of the annual performance review process, you are to rate each of your employees on a scale from 1-10. The following rating scale has been provided to you for assessing your employees:



[Page Break]

The firm believes supervisors should provide accurate ratings as doing so provides important to employees about their performance. Therefore, assume after you complete the evaluations, the next step will be to meet with a group of 4 other supervisors in your department. This committee will review the ratings of each employee and calibrate ratings across supervisors. **The goal of this committee is to reduce rating differences across the supervisors.**

[PCC_HIGH_SC Condition]
Here are the features of the calibration committee:

- The other committee members all manage teams in your department
- These supervisors have similar experience to your own
- All the teams of the supervisors in the committee are similar to your own
- All supervisors present their ratings to the committee for review
- You frequently interact and work with the other supervisors on the committee
- You care if the other committee members think you are a good supervisor

[Page Break]

[Scenario for PCC_LOW_SC Condition]
Here are the features of the calibration committee:

- All the other supervisors on your committee are fairly new to the company
- You do not have a lot of interaction with the other supervisors on your committee
- All the other supervisors are junior to you
- All supervisors present their ratings to the committee for review
- You do not care how your ratings compare to the other committee members

[Page Break]

Assume once this committee has finalized your evaluations, you will be responsible for discussing them with your employees. You would then submit your employees' evaluations to HR for inclusion in each employee's files. A summary of your team's ratings would also be sent to your supervisor as part of his/her resources for assessing your own performance. **As the firm believes how an employee performs reflects both the employee's ability and effort and the supervisor's ability to bring out the best in their employees.**

Panel C – Pilot Test One Results

Demographic Information

	N	Mean	Std Dev
Gender (Male=0)	207	0.454	0.499
Age	207	39.188	11.133
ProfCert (N= 0)	207	0.275	0.448
CCExp (N=0)	203	0.586	0.494
#EmpSup	207	31.275	91.181

Summary Statistics for Composite Social Comparison and Impression Management Score (SCIM)

	Median	**Minimum**	**Std Dev**	**25th Percentile**	**Mean**	**75th Percentile**	**Maximum**
NoPCC	4.000	1.400	1.498	2.600	3.948	5.200	6.400
PCC_SC_High	4.800	1.000	1.181	4.200	4.676	5.600	6.800
PCC_SC)_Low	3.600	1.000	1.449	2.600	3.749	4.600	7.000

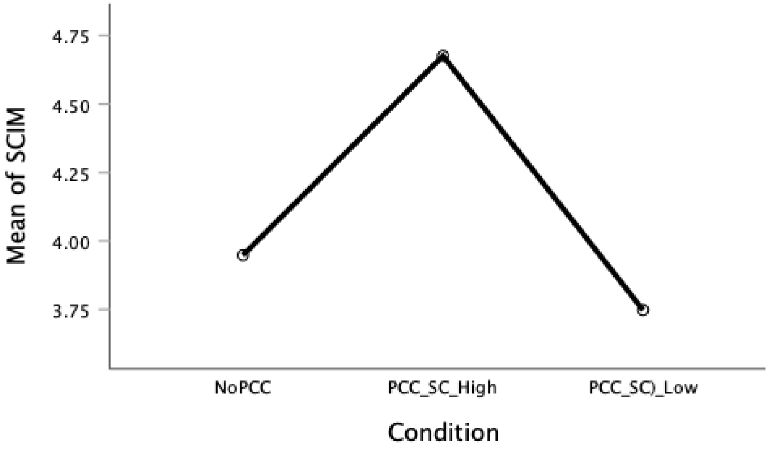
Note. The above table provides the means for the composite SCIM measure. SCIM is an average of the five main questions intended to capture participants' state of concern regarding social comparison and impression management desires (

Test of Means

(I) Condition	**(J) Condition**	**Mean Difference (I-J)**	**Std. Error**	**Sig.**
NoPCC	PCC_SC_High	-0.728	0.236	0.002
PCC_SC_High	PCC_SC)_Low	0.927	0.232	<0.001
NoPCC	PCC_SC)_Low	0.200	0.237	0.399

Note. Results are based on two-sided tests assuming equal variances.

Graph of SCIM Means by Condition



Note: SCIM is calculated based on Questions 1-5 Appendix E - Panel A. Factor analysis confirmed that these five questions reduce to one dimension (Eigenvalue of 3.00, explaining 59.993% of the variance).

Appendix G - Pilot Test Two

Panel A – Pilot Test Two Discussion

Overview

My second pilot test investigates how the subordinate profiles' presentation affects participants' ratings. I seek to identify a subordinate profile presentation that: 1) allows for enough rating variation between participants and 2) does not introduce random variation (i.e., noise) in responses. By doing so, I intend to identify a design that does not constrict my ability to observe support for my hypotheses, should that support exist.

I conduct a 3 x 1 x 10 mixed-design experiment to test how subordinate profile presentation affects participants' ratings. I test three subordinate profile presentation conditions to assess how the presentation of the information might impact ratings provided by participants. Each participant also evaluates ten subordinates using their assigned profile presentation condition.

Subordinate Profile Presentation Format

I test my subordinate profile's presentation format and included information as I designed these subordinate profiles specifically for this thesis study. My subordinate profiles need to provide enough details for participants to create enough realism regarding the performance review process, but also mitigate unintended variability in ratings. The outline of the subordinate profiles is adapted from the employee rating document presented in the field study conducted by Bol & Smith (2011). However, I created the skill categories, descriptive wording, rating scales, and other information newly for this experiment. I also use this pilot test to examine the use of the blue bars for indicating the preliminary ratings, as this is a new design feature to this type of experiment as well.

When testing my three designs I seek to select the presentation that 1) has profiles ratings with a similar sorting into performance categories as my designed distribution (i.e., low, moderate, and high performers), 2) a presentation that provides enough variation in ratings between participants such that design features have not limited my ability to find results (such as the 'blue bar' preliminary ratings), and 3) a presentation that does not provide too much variation in ratings between participants such that there is too much noise in participant ratings.

Task Details

All three subordinate profile presentations show the same 'preliminary assessment' of a subordinate's performance for each category (represented by a blue bar that spans three possible ratings; see example in Panel B, C and D).[58] I vary the presentation of the rating assessment to create my three conditions. In Condition 1, participants receive descriptive information (e.g., *"work is done by deadlines, but needs some improvement on general project time management"*) for each performance criteria category. Then, a single summary rating for each subordinate is requested. In Condition 2, participants do not receive descriptive information, but the request is still for a single summary rating for each subordinate. Finally, in Condition 3, participants are provided descriptive information for each performance criteria category (same as Condition 1). However, instead of a single summary rating, participants give a rating for each performance criteria category (i.e., six ratings for each profile). This three-condition design allows me to test whether descriptive information influences provided ratings and whether asking participants to provide a summary rating versus categorical ratings influences the overall subordinate rating.

[58] Subordinate profiles do not vary by condition in the six categories presented nor the "preliminary assessments" provided (i.e., the blue bars have the same placement in each condition).

Participants

I conduct Pilot Study Two with a group of MTurk workers as participants. I use an available filter on MTurk to exclude MTurk workers from participating in both Pilot Study One and Pilot Study Two. I pre-screen participants for a minimum of an undergraduate (associate or bachelor's) degree, experience supervising subordinates (to ensure that participants have the necessary work experience), and a 95% HIT approval rating (to increase the likelihood that participants complete the task effortfully). Two hundred and twenty-five MTurk workers participated in this pilot test. I removed four participants from the analysis due to their poor-quality responses.[59]

Pilot Test Two Results and Conclusion

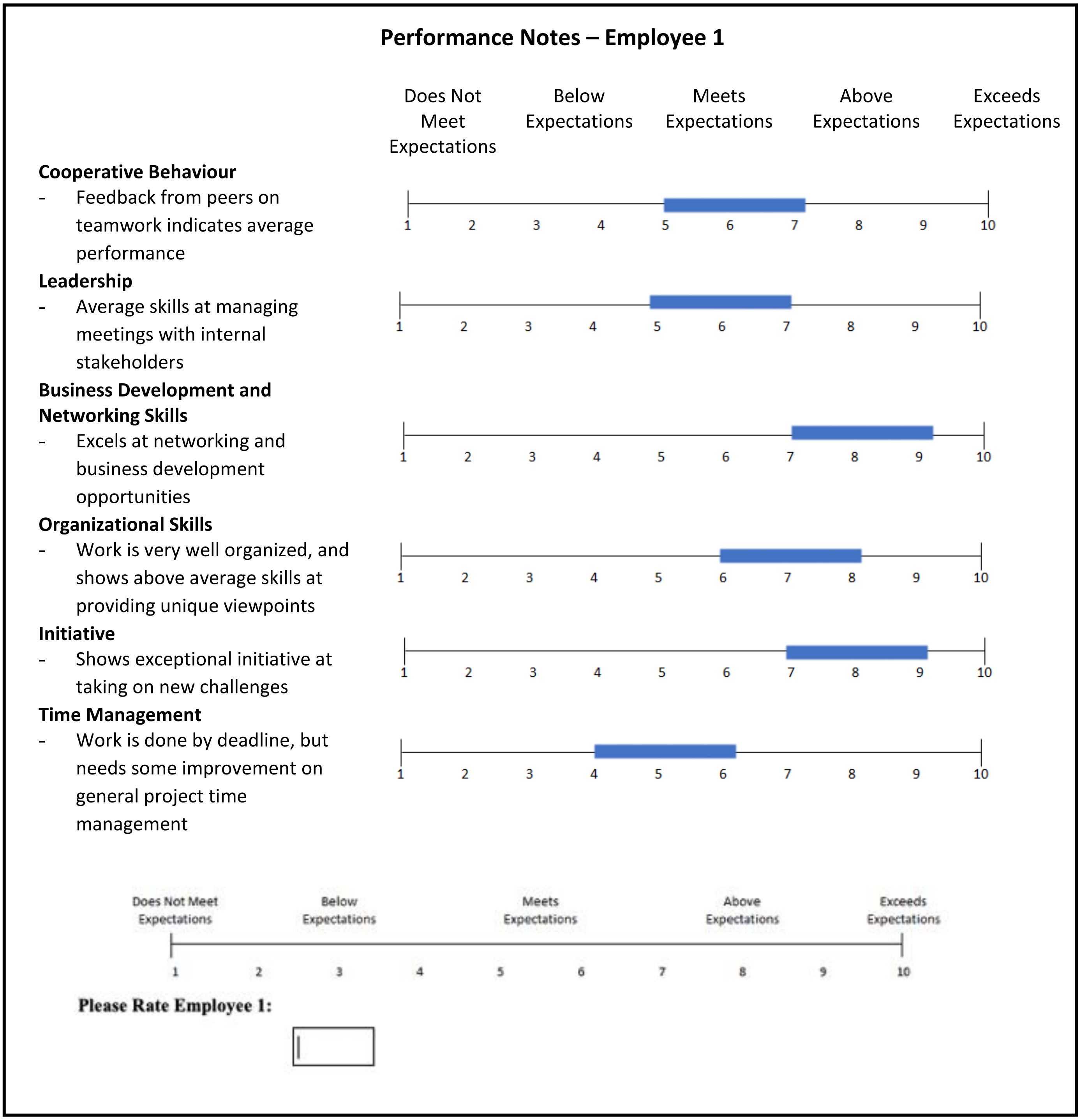
Results (Panel E) show that there is significant variation observed between the ten subordinate ratings ($p < 0.001$, two-tailed; Panel E). However, there is no significant difference in mean rating among the three subordinate profile presentation conditions ($p = 0.119$, two-tailed; Panel E).[60]

Overall, the subordinate profile presentation does not seem to significantly impact the mean ratings assigned to subordinates. Nor does the subordinate profile present seem to vary the distribution of ratings too narrowly or broadly. Therefore, I have selected Condition 2 to test my hypotheses since it offers the most straightforward presentation. Specifically, I ask participants to provide one summary rating per subordinate, and I do not include extra descriptive information for the performance criteria categories (Panel B).

[59] These four participants were removed because they provided the same rating to every subordinate.

[60] Analysis examining each subordinate rating as an individual measure produced similar results

Panel B - Example Subordinate Profile Condition 1



Performance Notes – Employee 1

Cooperative Behaviour

- Feedback from peers on teamwork indicates average performance

Leadership

- Average skills at managing meetings with internal stakeholders

Business Development and Networking Skills

- Excels at networking and business development opportunities

Organizational Skills

- Work is very well organized, and shows above average skills at providing unique viewpoints

Initiative

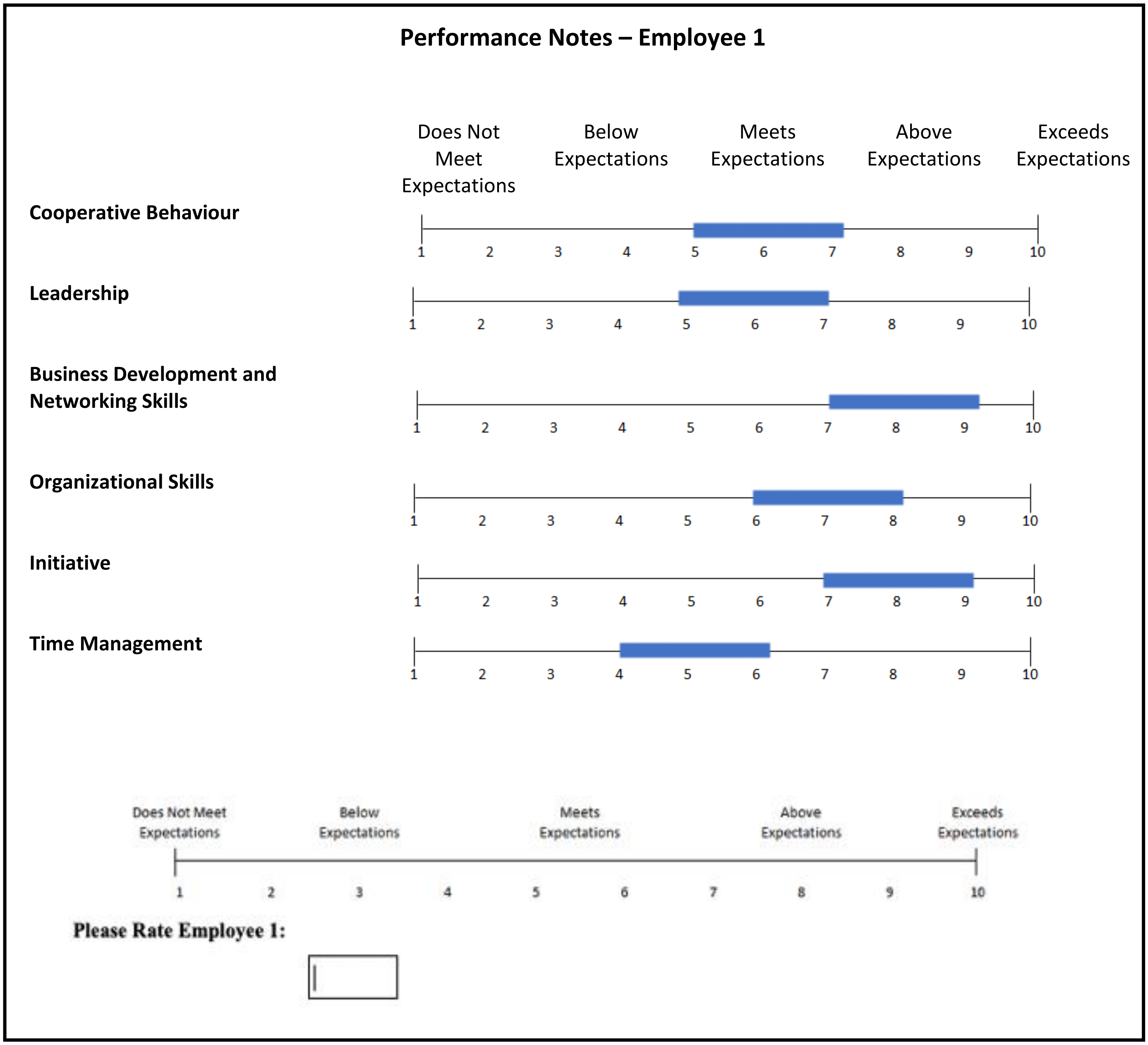
- Shows exceptional initiative at taking on new challenges

Time Management

- Work is done by deadline, but needs some improvement on general project time management

Please Rate Employee 1:

Panel C - Example Subordinate Profile Condition 2



Performance Notes – Employee 1

Cooperative Behaviour

Leadership

Business Development and Networking Skills

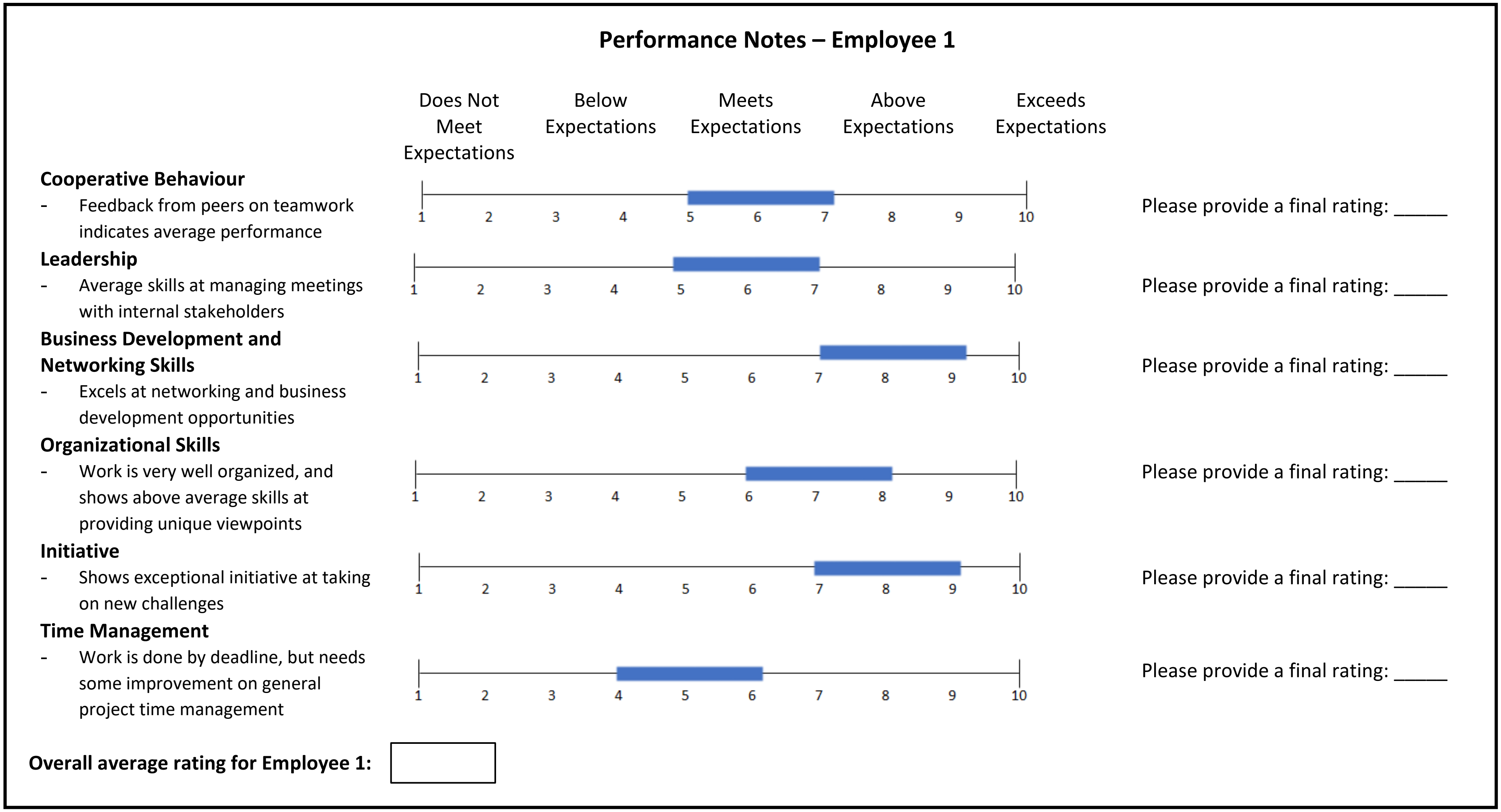
Organizational Skills

Initiative

Time Management

Please Rate Employee 1:

Panel D - Example Subordinate Profile Condition 3



Performance Notes – Employee 1

Does Not Meet Expectations | Below Expectations | Meets Expectations | Above Expectations | Exceeds Expectations

Cooperative Behaviour
- Feedback from peers on teamwork indicates average performance

Please provide a final rating: ______

Leadership
- Average skills at managing meetings with internal stakeholders

Please provide a final rating: ______

Business Development and Networking Skills
- Excels at networking and business development opportunities

Please provide a final rating: ______

Organizational Skills
- Work is very well organized, and shows above average skills at providing unique viewpoints

Please provide a final rating: ______

Initiative
- Shows exceptional initiative at taking on new challenges

Please provide a final rating: ______

Time Management
- Work is done by deadline, but needs some improvement on general project time management

Please provide a final rating: ______

Overall average rating for Employee 1: □

Note. In this condition, participants rate the subordinate for each category, and the overall rating is calculated and displayed by the system as an average of the six provided ratings.

Panel E – Pilot Test Two Results

Demographic Information

	N	Mean	Std Dev
Gender (Male=0)	221	0.335	0.473
Age	221	38.276	10.768
Professional Certification (No=0)	221	0.326	0.470
PCC Experience (N=0)	221	0.561	0.497
Number of Subordinate's Supervised	219	29.594	77.659

Descriptive Statistics by Condition for Each Employee Profile

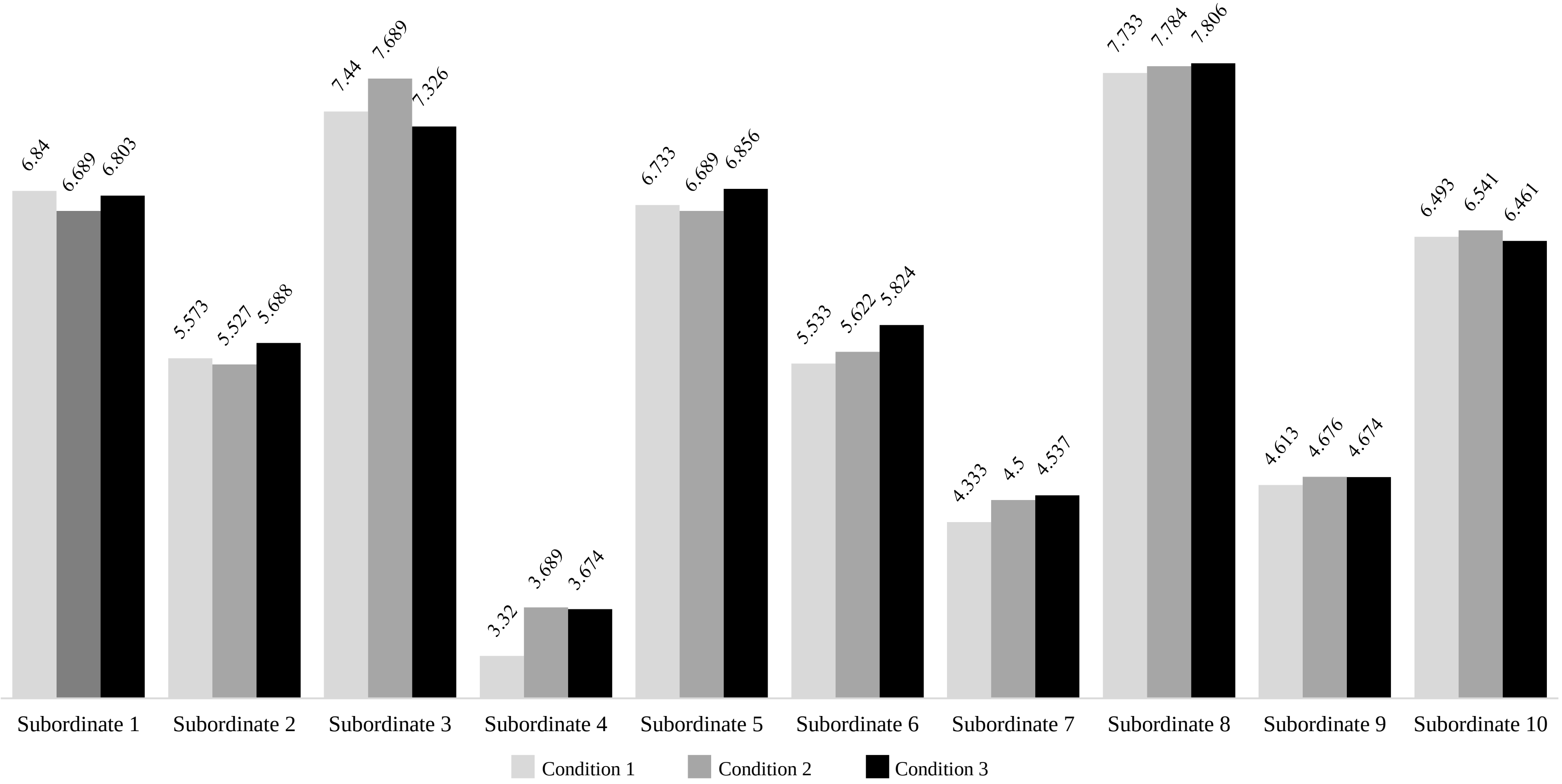
		Mean	Median	Mode	Minimum	Maximum	Std Dev
Subordinate 1	Condition 1	6.840	7.000	7.000	5.000	9.000	0.789
	Condition 2	6.689	7.000	7.000	5.000	9.000	0.739
	Condition 3	6.803	6.750	6.667	2.833	8.500	0.645
Subordinate 2	Condition 1	5.573	5.000	5.000	4.000	8.000	0.903
	Condition 2	5.527	6.000	6.000	4.000	8.000	0.798
	Condition 3	5.688	5.500	5.500	4.167	8.500	0.658
Subordinate 3	Condition 1	7.440	8.000	8.000	5.000	9.000	0.826
	Condition 2	7.689	8.000	8.000	5.000	10.000	0.757
	Condition 3	7.326	7.333	7.333	3.000	8.333	0.669
Subordinate 4	Condition 1	3.320	3.000	3.000	2.000	9.000	1.055
	Condition 2	3.689	3.000	3.000	2.000	8.000	1.238
	Condition 3	3.674	3.333	3.333	2.000	8.167	1.250
Subordinate 5	Condition 1	6.733	7.000	7.000	5.000	9.000	0.859
	Condition 2	6.689	7.000	7.000	3.000	9.000	0.905
	Condition 3	6.856	6.833	6.667	2.500	8.500	0.708
Subordinate 6	Condition 1	5.533	6.000	6.000	3.000	10.000	0.991
	Condition 2	5.622	5.000	5.000	3.000	9.000	1.003
	Condition 3	5.824	5.667	5.667	4.000	8.333	0.581
Subordinate 7	Condition 1	4.333	4.000	4.000	3.000	9.000	1.178
	Condition 2	4.500	4.000	4.000	2.000	8.000	0.969
	Condition 3	4.537	4.167	4.167	3.333	8.000	0.913
Subordinate 8	Condition 1	7.733	8.000	8.000	5.000	10.000	0.827
	Condition 2	7.784	8.000	8.000	3.000	10.000	1.101
	Condition 3	7.806	7.750	7.667	4.000	8.833	0.659
Subordinate 9	Condition 1	4.613	4.000	4.000	3.000	9.000	1.161
	Condition 2	4.676	5.000	5.000	2.000	10.000	1.061
	Condition 3	4.674	4.333	4.333	3.333	8.000	0.881
Subordinate 10	Condition 1	6.493	6.000	6.000	5.000	9.000	0.828
	Condition 2	6.541	7.000	7.000	5.000	9.000	0.725
	Condition 3	6.461	6.333	6.333	3.500	8.500	0.650

Notes. The mean ratings and standard deviation of the ratings are presented above by condition for each subordinate.

Ratings for each subordinate are provided based on a scale ranging numerically from 1-10 and labelled from "*Did Not Meet Expectations*" to "*Exceeds Expectations*" See Appendix G – Panel B, C or D for example, subordinate profiles showing the exact scale used.

Condition 1 - N = 75 - Participants assess overall rating with descriptive information.
Condition 2 - N = 74 - Participants assess overall rating without descriptive information
Condition 3 - N = 72 - Participants assess individual category ratings with descriptive information.

Graph of Estimated Marginal Means of Subordinate Ratings by Condition



Notes. Mean ratings for each subordinate profile by condition
Condition 1 – With Descriptions
Condition 2 – No Descriptions
Condition 3 – With Descriptions and Rating done for each Category

General Linear Model Repeated Measures Test

Mauchly's Test of Sphericity

					Epsilon
Within Subjects Effect	**Mauchly's W**	**Approx. Chi-Square**	**df**	**Sig.**	**Greenhouse-Geisser**
Subordinate	0.049	647.405	44	<0.001	0.447

Tests of Within-Subjects Effects

	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Subordinate	Greenhouse-Geisser	3787.666	4.027	940.657	773.136	<0.001
Subordinate * Condition	Greenhouse-Geisser	15.822	8.053	1.965	1.615	0.116
Error (Subordinate)	Greenhouse-Geisser	1068.002	877.803	1.217		

Notes. Tested for differences between conditions using employee ratings and a within-subjects repeated measure. As Mauchly's test of Sphericity was significant, a Greenhouse-Geisser transformation was applied.

The alternative Huynh-Feldt transformations also showed similar results.

**** p<0.01, ** p<0.05, * p<0.1*

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	77488.530	1	77488.530	24450.032	<0.001
Condition	4.334	2	2.167	0.684	0.506
Error	690.899	218	3.169		

Appendix H – Subordinate Midpoint Ratings

Categorization of Performance Level	Subordinate Profile Presented to Participant	Median of the Midpoints of Provided Ranges	Mean of the Midpoints of Provided Ranges
Low Performers			
Low_Performer_1	Subordinate Profile 4	3.000	3.333
Low_Performer_2	Subordinate Profile 7	4.000	4.167
Low_Performer_3	Subordinate Profile 9	4.000	4.333
Moderate Performers			
Moderate_Performer_1	Subordinate Profile 1	6.500	6.667
Moderate_Performer_2	Subordinate Profile 2	5.500	5.500
Moderate_Performer_3	Subordinate Profile 6	6.500	5.667
Moderate_Performer_4	Subordinate Profile 10	6.500	6.333
High Performers			
High_Performer_1	Subordinate Profile 3	7.500	7.333
High_Performer_2	Subordinate Profile 5	7.000	6.667
High_Performer_3	Subordinate Profile 8	8.000	7.667

Note. This table contains the medians and means of the midpoints of preliminary rating ranges provided to participants for each subordinate profile. The median is taken to form the *Subordinate Midpoint Ratings* for use in determining the subordinate performance levels and my measure of *Leniency Bias*.

Appendix I - Variable Names and Definitions

Variable Name	**Definition**
PCC	Dummy Variable for Absence (0) or Presence (1) of a PCC
RDG	Dummy Variable for Absence (0) or Presence (1) of RDG
Subordinate Midpoint Rating	To assist with calculating a Leniency Bias measure, the median of the midpoints is taken as the non-biased ratings of that subordinate profile.
Low Performers	Based on the *Subordinate Midpoint Rating,* subordinate profiles with median-based midpoints below 5 are classified as *Low Performers.* These subordinate profiles are: Subordinate Profile 4 Subordinate Profile 7 Subordinate Profile 9
Moderate Performers	Based on the *Subordinate Midpoint Rating,* subordinate profiles with median-based midpoints above 5 and below 7 are classified as *Moderate Performers.* These subordinate profiles are: Subordinate Profile 1 Subordinate Profile 2 Subordinate Profile 6 Subordinate Profile 10
High Performers	Based on the *Subordinate Midpoint Rating,* subordinate profiles with median-based midpoints 7 and above are classified as *High Performers*. These subordinate profiles are: Subordinate Profile 3 Subordinate Profile 5 Subordinate Profile 8
Low_Performer_[1-3]	A variable that relabels the subordinate profiles based on their performance level classification. Relabelling is as follows: Subordinate Profile 4 → Low_Performer 1 Subordinate Profile 7 → Low_Performer 2 Subordinate Profile 9 → Low_Performer 3
Moderate_Performer_[1-4]	A variable that relabels the subordinate profiles based on their performance level classification. Relabelling is as follows: Subordinate Profile 1 → Moderate_Performer 1 Subordinate Profile 2 → Moderate_Performer 2 Subordinate Profile 6 → Moderate_Performer 3 Subordinate Profile 10 → Moderate_Performer 4

High_Performer_[1-3]	A variable that relabels the subordinate profiles based on their performance level classification. Relabelling is as follows: Subordinate Profile 3 → High_Performer 1 Subordinate Profile 5 → High_Performer 2 Subordinate Profile 8 → High_Performer 3
Subordinate Midpoint Rating	To assist with calculating a Leniency Bias measure, the median of the midpoints is taken as the non-biased ratings of that subordinate profile (*Subordinate Midpoint Rating*).
Leniency Bias Measure	In calculating the Leniency Bias measure, the median of the midpoints is taken as the non-biased ratings of that subordinate profile (*Subordinate Midpoint Rating*). Therefore, to evaluate the bias displayed, this *Subordinate Midpoint Rating* is subtracted from the rating given by the participants to create a measure of bias shown by that participant. (See Table 1 - Panel D for the midpoints of each subordinate profile)
Leniency_Low _[1-3]	Leniency Bias Measure for Low_Performer_[1-3]
Leniency_Moderate _[1-4]	Leniency Bias Measure for Moderate_Performer_[1-4]
Leniency_High_[1-3]	Leniency Bias Measure for High_Performer_[1-3]
Overall Average	Overall average of all 10 subordinates rated by a participant
Evaluation Experience	Participant reported they had experience evaluating subordinates
CC Experience	Participant reported they had experience participating in a Calibration Committee
Gender	Gender of Participant (Male = 0)
Age	Age of Participant
Time (min)	Time to complete the study in minutes
SC_Think	Response to Post Experimental Question: *"To what extent did you think about your employees' ratings relative to other supervisors' employee ratings when completing your evaluations?"*
SC_Concern	Response to Post Experimental Question: *"To what extent were you concerned about how your employee ratings compared to other supervisors' employee ratings when you rated your employees?"*
SC_ Interferes	Response to Post Experimental Question: *"To what extent did thinking about comparing your ratings with other supervisors' ratings interfere with your ability to concentrate when you rated your employees?"*
IM_DemAbility	Response to Post Experimental Question: *"To what extent did you consider your employee ratings as being an opportunity to demonstrate your ability as a supervisor to your peers when you rated your employees?"*
IM_PeerOpin	Response to Post Experimental Question:

	“To what extent did you consider other supervisors’ opinions about your abilities as a supervisor when you rated your employees?”
IM_SubFairPercep	Response to Post Experimental Question: *“To what extent did you consider how fair your employees would think their ratings were when you rated your employees?”*
IM_SupOpin	Response to Post Experimental Question: *“To what extent did you consider your own supervisor’s opinion of your ability as a supervisor when you rated your employees?”*
PCC_AdjDown	Response to Post Experimental Question: *“To what extent did you consider how the calibration committee might adjust your ratings downwards when rating your employees?”* Note: Only asked when PCC Present
PCC_Negotiate	Response to Post Experimental Question: *“To what extent did you consider whether you might have to negotiate your ratings in the calibration committee when rating your employees?”* Note: Only asked when PCC Present
N_Fairness	Response to Post Experimental Question: *“To what extent do you believe a manager should consider the fairness of the ratings when evaluating employees?”*
N_High	Response to Post Experimental Question: *“To what extent do you believe a manager should NOT provide too many high ratings when evaluating a group of employees?”*
N_Low	Response to Post Experimental Question: *“To what extent do you believe a manager should NOT provide too many low ratings when evaluating a group of employees?”*
N_ConsidPeerOp	Response to Post Experimental Question: *“To what extent do you believe a manager should consider how their peers will perceive the ratings they provide?”*
N_Dist	Response to Post Experimental Question: *“To what extent do you believe a manager should provide a wide distribution of ratings when evaluating employees?”*
N_Accuracy	Response to Post Experimental Question: *“To what extent do you believe a manager should provide the most accurate ratings possible for employees?”*