RICE UNIVERSITY

Why do raters pursue different rating goals? The role of rater personality and accountability context

by

Jisoo Ock

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

[Signatures of committee members]

Frederick L. Oswald, Chair
Professor, Psychology

Margaret E. Beier
Associate Professor, Psychology

Anton J. Villado
Assistant Professor, Psychology

D. Brent Smith
Associate Professor, Management & Psychology

HOUSTON, TEXAS
DECEMBER, 2014
ABSTRACT

Why Do Raters Pursue Different Rating Goals?
The Role of Rater Personality and Accountability Context

By
Jisoo Ock

Although performance ratings are often put to use in making critical managerial decisions that have important practical implications for organizations (e.g., personnel decisions), there is a wide concern among both researchers and practitioners alike that performance ratings are generally poor indicators of the job performance construct. Among the different aspects of the criterion problem, performance appraisal literature has emphasized the role of rater motivation in performance appraisal, which is determined in part by the dispositional characteristics of the raters, and their interaction with the context in which performance appraisal takes place, as an important factor that determines the quality of performance ratings. The goal of the current study was to examine the effect that rater personality (agreeableness, conscientiousness, neuroticism), and accountability context (ratee, supervisor, baseline) has on performance ratings through the interactive effect that rater personality and accountability context has on raters’ intentions in the performance appraisal process.
ACKNOWLEDGEMENTS

Throughout my tenure as a graduate student, I have been blessed with an incredible opportunity to study and conduct research on topics in I/O psychology, and even more incredible supporters who have made completion of this paper possible. First and foremost, I want to thank my advisor, Fred Oswald, for his guidance and support he has given me throughout my graduate school career. Fred, I really cannot thank you enough for everything you have done for me over the years, and I could not have asked for a better advisor. Thank you.

I would also like to thank Margaret Beier, Brent Smith, and Anton Villado for serving on my dissertation committee, and providing insightful comments and feedback about my work. I know that this paper improved so much from your help, and I would like to express my most sincere gratitude for all the time and effort you have put into serving as a committee member.

I would also like to sincerely thank Mikki Hebl. One of the things that I will regret most about my graduate school career is not working with you sooner. I have utmost respect for you as a researcher, a teacher, and more importantly, as a person. Thank you so much for your support and guidance over the years.

This paper would not be complete without the support of my girlfriend (and soon-to-be-wife), Gayoung Cho. Meeting you was the best thing that has ever happened to me. I am so thankful for your support and love throughout my years in graduate school, and I am so glad that I will always have you along my side as we continue our journey in life together.

I would also like to thank my family and friends for always being there to celebrate the good times together, and console me during the bad ones (and there has
been a quite of few of them). I could not have accomplished what I have accomplished without your support.

Finally, I would like to thank my parents, Wonho Ock and Hyeran Suh, for everything they have done for me. Mom and dad, both of you have been my biggest role models in life, and everything I have accomplished and will accomplish in the future will be because of your love, support, and sacrifices. I dedicate this dissertation to both of you. Thank you, and I love you both so much.
TABLE OF CONTENTS

ABSTRACT .......................................................................................................................... ii
ACKNOWLEDGEMENTS ................................................................................................. iii
LIST OF TABLES ................................................................................................................ vi
LIST OF FIGURES .............................................................................................................. vii
CHAPTER 1: INTRODUCTION ........................................................................................ 1
CHAPTER 2: LITERATURE REVIEW ............................................................................... 4
  2.1 Effect of Social Context on Performance Appraisal ............................................... 4
  2.2 Effect of Rating Context and Rater Characteristics on Performance Appraisal .... 8
  2.3 Person-Situation Interaction in Performance Appraisal .......................................... 16
CHAPTER 3: STUDY OVERVIEW .................................................................................... 18
  3.1 Study Hypotheses ...................................................................................................... 22
CHAPTER 4: METHOD ..................................................................................................... 30
  4.1 Sample ....................................................................................................................... 30
  4.2 Procedure and Experimental Manipulation ............................................................. 33
  4.3 Materials .................................................................................................................... 34
  4.4 Measures ................................................................................................................... 34
  4.5 Dependent Variables ............................................................................................... 36
  4.6 Analysis ...................................................................................................................... 36
CHAPTER 5: RESULTS .................................................................................................... 41
  5.1 Perceived Accountability ......................................................................................... 41
  5.2 Mediation and Moderated Mediation Analysis ....................................................... 41
  5.3 Mean Overall Performance Rating and Rating Accuracy Comparisons ............. 62
CHAPTER 6: DISCUSSION .............................................................................................. 72
  6.1 Theoretical Implications ......................................................................................... 78
  6.2 Practical Implications .............................................................................................. 79
  6.3 Limitations and Future Research ............................................................................ 81
  6.4 Conclusion ................................................................................................................ 84
REFERENCES ..................................................................................................................... 85
APPENDIX: Rating Goal Questionnaire ......................................................................... 94
List of Tables

Table 1      Participant Demographics ................................................................. 32
Table 2      Means, Standard Deviations, and Correlations between Variables .......... 43
Table 3a     Simple Mediation Analysis Results for Rater Agreeableness ................. 45
Table 3b     Moderated Mediation Analysis Results for Rater Agreeableness .......... 47
Table 4a     Simple Mediation Analysis Results for Rater Neuroticism .................... 51
Table 4b     Moderated Mediation Analysis Results for Rater Neuroticism ................ 53
Table 5a     Simple Mediation Analysis Results for Rater Conscientiousness .......... 57
Table 5b     Moderated Mediation Analysis Results for Rater Conscientiousness ....... 59
Table 6      Mean Difference between Personality and Accountability Conditions .... 64
Table 7      Rater Agreeableness × Accountability Context ANOVA Results ............ 66
Table 8      Rater Neuroticism × Accountability Context ANOVA Results ............... 68
Table 9      Rater Conscientiousness × Accountability Context ANOVA Results ....... 71
List of Figures

Figure 1    Mediated Model of the Performance – Performance Rating Relationship ........ 15
Figure 2a   Mediating Effect of Rating Goals on the Relationship between Rater Agreeableness and Overall Performance Rating ......................................................... 48
Figure 2b   Mediating Effect of Rating Goals on the Relationship between Rater Agreeableness and Rating Accuracy ................................................................. 49
Figure 3a   Mediating Effect of Rating Goals on the Relationship between Rater Neuroticism and Overall Performance Rating .................................................. 54
Figure 3b   Mediating Effect of Rating Goals on the Relationship between Rater Neuroticism and Rating Accuracy ................................................................. 55
Figure 4a   Mediating Effect of Rating Goals on the Relationship between Rater Conscientiousness and Overall Performance Rating ........................................ 60
Figure 4b   Mediating Effect of Rating Goals on the Relationship between Rater Conscientiousness and Rating Accuracy ......................................................... 61
Chapter 1
Introduction

Organizations have a clear and common need to assess their employees’ level of job performance accurately, because this informs important decisions that have major implications for organizational operations (e.g., high-stakes personnel decision, identifying employee development needs). For example, if supervisors base their evaluation of employees’ performance on irrelevant factors, such as how much they like them as opposed to their actual job performance (Wayne & Liden, 1995), then it is foreseeable that this will result in several errors in personnel decisions (i.e., promoting employees who should not be promoted; failing to promote employees who should be promoted), the consequences of which may be dire for the future of an organization (e.g., talented employees leaving to seek better opportunities; organizational performance suffering).

Despite the fact that performance ratings are often put to use in making such critical decisions that have important practical implications for organizations, there is a wide concern that performance ratings are typically poor representations of employees’ actual job performance (Cascio, 1978; Landy & Farr, 1980, 1983; Murphy & Cleveland, 1991, 1995). What is more problematic is that rating error seems to occur often in practice. For example, it has been reported that 80% of all employees who are evaluated on job performance on a 7-point Likert-type scale are rated as being highly effective (6 or 7). Although such skewed distribution is certainly possible (especially given that organizations can dismiss poor performers), the fact that it is so common across organizations (Bretz, Milkovich, & Read, 1992) has led researchers and practitioners to question the validity of performance ratings. Additionally, as evidenced by the dismal average interrater correlation for job
performance ($\rho = .52$; Viswesvaran, Ones, & Schmidt, 1996), there clearly seem to be reasons to doubt the trustworthiness of performance ratings (however, see LeBreton, Burgess, Kaiser, Atchley, & James, 2003, for evidence that job performance interrater reliability is negatively biased because of range restriction). Some researchers have taken the argument so far as suggesting that organizations should do away with performance appraisal entirely (Coen & Jenkins, 2000), and some have called into question the validity of performance criterion based validation research (Osterman, 2007). Poor quality of performance ratings is especially concerning given the wide use of formal performance appraisal system in organizations (Bernthal, Sumlin, Davis, & Rogers, 1997), and the significance of the impact that inaccurate performance ratings can have on organizations.

However, considering the context in which performance appraisal typically takes place, it is understandable why performance ratings tend to be poor indicators of employees’ job performance. In many organizations, employees are often held responsible for evaluating other employees with whom they work together (and must continue to work together after the appraisal in many cases). In these situations, performance appraisal is understandably difficult because raters are directly responsible for determining the outcomes of performance appraisal, which are often tied to real consequences for the ratees that they evaluate (e.g., promotion, remuneration). Thus, raters must approach performance appraisal carefully with considerations toward a host of practical and interpersonal factors that are in play in the social environment of the organization. For example, supervisors who are concerned about damaging the interpersonal relationships with their employees might prefer the benefits of providing lenient ratings (e.g., preserve or enhance supervisor—subordinate relationship) to the costs associated with providing such inaccurate
ratings (e.g., employees being rewarded despite inadequate performance). Although performance appraisal with considerations toward its implications for social environment of the organization might have its own benefits (e.g., high quality leader-member exchange is related to more organizational citizenship behaviors; Setton, Bennett, & Liden, 1996), it definitely undermines the accuracy of performance ratings.

The key factor that complicates performance appraisal for raters is that they are held responsible for whatever ratings that they provide, or whatever organizational decisions that result from the ratings that they provide, all of which could have important implications for the raters, the ratees, or the organization. Then, the validity of performance ratings should be influenced by how raters choose to account for their ratings, which should vary considerably depending on the context of the performance appraisal, as well as raters’ individual difference characteristics that influence how they behave in different performance appraisal contexts.

Based on the theoretical proposition that performance ratings are largely determined by raters’ intentions in the performance appraisal process (Murphy & Cleveland, 1991, 1995), the goal of this study was to examine the effect that rating audience and individual difference in raters’ personality traits have on how raters choose to approach performance appraisal, and how these factors influence the mean level, and accuracy of the ratings that raters provide.
CHAPTER 2

Literature Review

As evidenced by the long struggle with the *criterion problem* in personnel psychology (Austin & Villanova, 1992), performance appraisal has been recognized as one of the most challenging issues in the field of organizational psychology for both researchers and practitioners alike. This section highlights some of the key aspects of previous literature on performance appraisal that are relevant to this study. Namely, I will discuss the goal-directed performance appraisal perspective and its relevance to empirical studies that have demonstrated significant effects of rater personality and accountability contexts on performance ratings. Additionally, I will discuss Tett and Burnett’s (2003) trait activation theory as the theoretical basis for explaining my study hypotheses regarding the joint effect of accountability context and rater personality on performance ratings.

Effect of Social Context on Performance Appraisal

Most organizations are social by nature. In an organization, its members not only work together on job tasks, but also often develop interpersonal relationships beyond the work context. The social aspect of organizations makes performance appraisal a daunting challenge for raters as they are tasked with an obligation to accurately and objectively evaluate employee’s job performance (which might include criticizing or even punishing employees), but at the same time, are obliged to consider the potential consequences of how their evaluations might influence their personal relationships with the ratees (e.g., upsetting employees by providing ratings that are too much lower than the highest ranking), or how management might judge their ability (e.g., rating employees too high might give the impression that the rater is too lenient; rating employees too low might give the impression that the rater is an
incompetent manager) as a function of the ratings that they provide. Given the complicated context in which performance appraisal typically takes place, researchers suggested that raters are motivated to successfully navigate through the complexities involved in the performance appraisal process (Cleveland & Murphy, 1992; Murphy & Cleveland, 1991, 1995); a perspective that is described as the goal-directed performance appraisal (Murphy & Cleveland, 1995).

**Goal-directed performance appraisal.** Goal-directed performance appraisal may be described as deliberate strategic decision on the part of the raters to conduct performance appraisal with regards to potential benefits or consequences of the ratings that they provide (Harris, 1994; Kane, 1994; Murphy & Cleveland, 1991, 1995). Impressionistic (direct rater reports of the nature and extent of rating errors) and anecdotal evidence (Bjerke, Cleveland, Morris, & Wilson, 1987; Gioia & Longenecker, 1994; Longenecker, Gioia, & Sims, 1987) suggests that raters indeed think about the consequences of their ratings, and that these considerations influence raters’ intentions in the appraisal process as well as the ratings that they provide. More importantly, research has shown that raters often pursue rating goals that do not necessarily align with the rating goals that the organization would like the raters to pursue in performance appraisal (Bjerke et al., 1987; Harter, 1986). For example, Longenecker et al. (1987) interviewed 60 business executives about their impression of how performance appraisal is conducted in practice. The interviewees’ testimonials suggested that raters intentionally distorted performance ratings on a regular basis, often prioritizing the considerations for benefits or consequences of their ratings for themselves or the ratees ahead of accurately evaluating the ratees’ performance. In fact, in a survey of raters, ratees, and administrators of performance appraisal, majority of respondents reported that deliberate rating distortion is much
more likely to be responsible for poor quality of performance ratings than raters’ inadvertent cognitive errors (Bernardin & Villanova, 1986).

Research has also shown that raters’ rating intentions are indeed important predictors of performance ratings. In one of the first studies that examined the relationship between rating goals and performance ratings, Murphy, Cleveland, Skattebo, and Kinney (2004) asked undergraduate students in five different courses to indicate their perception of importance of four types of rating goals (identifying weaknesses, identifying strengths, providing fair ratings, motivating ratees) once at the beginning of the semester, and once at the end of the semester (students were also asked to complete a standard university-wide instructor evaluation at the end of the semester). The authors examined the relationship between students’ perceived rating goal importance and instructor evaluation ratings within each classroom, and found a moderate to strong effect ($R^2$ ranging from .09 to .45, with a median of .25). Interestingly, the authors also found that the goal importance ratings collected at the beginning of the semester were only moderately correlated to the goal importance ratings collected at the end of the semester (average $r = .27$), and that the goal importance ratings collected at the end of the semester explained incremental variance in instructor evaluation ratings above and beyond the variance explained by the goal importance ratings collected at the beginning of the semester. Further analysis showed that students who rated the instructors favorably were more likely to rate “identifying strengths” goal as more important at the end of the semester than at the beginning of the semester. The instructors in Murphy et al.’s (2004) study were in fact effective instructors (each instructor’s average rating was above 4 on a 5-point scale). Thus, the results suggest that students became more motivated to emphasize rating goal that best supports their high ratings (identifying strengths) as students
observed a semester-long display of effective performance by the instructors. These findings can be interpreted as that not only does raters’ rating intentions predict performance ratings, raters actively modify their rating strategies as they gather more information about the ratee.

In a different study by Wong and Kwong (2007), the authors found that manipulating the raters’ rating intentions in performance appraisal influenced the mean level of ratings and the discriminability of the ratings (standard deviation of ratings). In this study, the authors instructed undergraduate students enrolled in a human resource management course to evaluate performance of their team members (14 teams, each with 7 or 8 members) four times, each time with one of four rating goals: identification goal (control condition; identifying strengths and weaknesses), harmony goal, fairness goal, and motivation goal (students completed these peer ratings once at the middle of the semester and once at the end of the semester). The authors found that relative to when the students were instructed to evaluate their team members with an identification goal, they provided significantly higher ratings when they were instructed to evaluate their team members with a motivating goal, significantly lower ratings when they were instructor to evaluate their team members with a fairness goal, and both higher and less discriminatory ratings when they were instructed to evaluate their team members with a harmony goal. In addition to Murphy et al.’s (2004) finding that rater intentions predict performance ratings, Wong and Kwong’s (2007) results showed that different rater intentions produce meaningful differences in the mean level and variance of the ratings that raters provide.

If rater intentions are important determinants of performance ratings, then it is important to understand what factors influence how raters choose to approach performance appraisal. As mentioned earlier, researchers have begun to emphasize
the importance multifaceted organizational contexts, and individual difference in rater characteristics that influence how raters respond to different organizational contexts as important factors that influence performance ratings (Harris, 1994; Levy & Williams, 2004). In the following section, I review the key findings on the effect of rating context and rater personality on performance ratings.

**Effect of Accountability and Rater Characteristics on Performance Appraisal**

**Accountability.** Accountability refers to “being answerable to audiences for performing up to certain prescribed standards, thereby fulfilling obligations, duties, expectations, and other charges” (Schlenker, Britt, Pennington, Murphy, & Doherty, 1994, p. 634). According to Tetlock and colleagues’ (e.g., Tetlock, 1985; Tetlock & Boettger, 1989; Tetlock, Skitka, & Boettger, 1989) extensive work on the effect of accountability on decision-making, decision-makers, when held accountable, typically account for their decision by choosing a solution that allows them to assume the least amount of liability and help them maintain approval from whoever the decision-makers must account their decision to. In other words, when decision-makers are held accountable, they tend to feel pressure to satisfy the audience to whom they are held accountable, which in turn prompts them to provide a solution that the audience would presumably approve. In the context of performance appraisal, accountability theory suggests that raters will scan their environment for cues regarding the type of ratings that would be easiest to justify, and provide ratings that follow that approach.

There are broadly two coping strategies that decision-makers pursue to account for their behavior, and whether they choose one strategy versus the other largely depends on whether or not their audience has a generally explicit preference regarding the decision-makers’ decision. Specifically, when the audience preference is obvious, decision-makers typically account for their behavior by pursuing a
decision that reflects the views of the audience (Tetlock, 1983; Tetlock et al., 1989). Under this strategy, called acceptability heuristic (Tetlock, 1985), decisions are easy to justify because they are specifically tailored to meet the prescribed preference that the audience has about the decisions.

Previous studies have shown that raters generally engage in acceptability heuristic, meaning that raters generally submit to the apparent demands of the audience. For example, raters who were led to believe that they would have to engage in a face-to-face feedback session with the ratee after the appraisal provided significantly higher ratings than raters who were not expecting to provide feedback (e.g., Fisher, 1979; Ilgen & Knowlton, 1980; Klimoski & Inks, 1990; Mero, Guidice, & Brownlee, 2007; Waung & Highhouse, 1997). Interestingly, it was also found in these studies that rating inflation was especially high when the ratee’s performance was relatively low, but not when it was too low. Klimoski and Inks (1990) concluded that because ratees typically would prefer positive evaluation of their own performance, raters held accountable to ratees engaged in rating inflation to please the ratees, and assume the least amount of liability in preparation for the anticipated face-to-face feedback session.

Similar results were found in studies that manipulated accountability to other audience groups. In addition to the ratees, raters might need to account for their ratings to managers overlooking the appraisal process (Harris, 1994; Mero et al., 2007). Accountability to a person of authority creates a context that is similar to Tetlock’s conception of acceptability heuristic. That is, authority figure has higher status than the rater, and this relative difference in authority should pressure raters to provide ratings that are consistent with the authority’s rating preference (Mero et al., 2007).
Previous studies have shown that raters generally submit to the demands of the higher authority. For example, raters who were told that they would have to discuss their evaluation of an employee with an expert, or an experimenter (a group that presumably expects accurate ratings) generally provided significantly more accurate and lower ratings (although accurate ratings do not necessarily imply lower ratings; Murphy & Balzer, 1989) than raters who were not held accountable (e.g., Curtis, Harvey, & Ravden, 2005; Jones, 1992). In fact, Mero and Motowidlo (1995) found that raters were significantly more lenient when they were informed that supervisors in the experiment preferred lenient ratings, whereas raters were significantly more stringent when they were informed that supervisors in the experiment preferred accurate ratings. In line with Tetlock’s conception of acceptability heuristic, these results clearly indicate that raters generally provide ratings that follow the preference of the audience when the audience’s preference is obvious.

However, when the audience expectations are not readily apparent, decision-makers account for their decision by developing a solution that is supported by a careful and comprehensive assessment of the quality of the decision (Schlenker, 1986; Tetlock, 1985). In other words, decision-makers account for their decision by developing a high quality decision that is justifiable from an objective standpoint so as to preempt any potential criticism from the audience. This approach to account for one’s decision is called preemptive self-criticism (Tetlock, 1985). The goal of this strategy is not necessarily to satisfy the audience. Rather, in this approach, decision-makers develop decisions that are defensible, even if they turn out to be in contrast to the expectations of the audience.

Previous studies that examined preemptive self-criticism in the context of performance appraisal has shown that raters tend to engage in careful and
comprehensive processing of ratee information when held accountable to an audience without a clear preference about the rater’s ratings, and such critical approach to performance appraisal produced more accurate ratings. For example, Mero and Motowidlo (1995) found that accountable raters who were not given any information about the audience’s preference were significantly more attentive to the ratee’s performance information, and took more notes about the ratee’s performance compared to raters who were not held accountable, or even raters who were held accountable but were also given information about the preference of the audience.

In line with the accountability theory, research findings on the effect of accountability on performance ratings summarized above generally indicate that accountability serves as an influential factor that motivates raters to choose a solution that is more easily defensible or easy to justify to the audience. However, how raters choose to account for their ratings vary depending on what the audience prefers, and whether the audience preferences are known to the raters or not, as these factors change what the “best” approach for making the ratings justifiable to the audience.

**Accountability manipulation.** Many researchers who examined the effect of accountability on performance ratings generally manipulated audience (e.g., ratee vs. supervisor vs. control), anticipation of future feedback session (e.g., feedback vs. no feedback), method of feedback (e.g., written vs. face-to-face), or identifiability (e.g., identifiable vs. anonymous) and assumed that accountability is the underlying explanatory variable for the effect of the manipulation on performance ratings (Roch & McNall, 2007). These manipulations, while theoretically linked to accountability, do not directly capture perceived accountability, making it unclear if the effect of the accountability manipulations on performance ratings is really driven by the raters’ feelings of accountability or some other demand characteristics in the experimental
setting. Previous studies have also suggested that performance appraisal literature should consider more carefully the relationship between accountability manipulations and raters’ feelings of accountability. For example, Roch, Ayman, Newhouse, and Harris (2005) examined the effect that audience and identifiability had on raters’ feelings of accountability, and found that the two manipulations did not directly influence accountability. In the current study, I directly asked participants about their feelings of accountability, and examined the relationship between the accountability manipulation and participants’ perceived accountability. Additionally, I examined the effect that accountability manipulation and the participants’ feelings of accountability had on the study results.

**Rater personality.** Evaluating employees’ job performance is an important and sensitive process that takes place in almost all organizations. As tough as performance appraisal is for many raters, how raters handle the challenges in the performance appraisal process is likely to vary depending on individual differences in rater personality that elicit systematically different behavioral tendencies in interacting with others (Hough, 1992).

Research on the effect of rater personality on performance ratings has primarily focused on two traits: agreeableness and conscientiousness. Kane, Bernardin, Villanova, and Peyrefitte (1995) proposed that because agreeable individuals tend to be kind, generous, and sympathetic towards others, agreeable raters would prefer to focus on the positive aspects of ratees’ performance and find it difficult to be critical, especially when there is anticipation for ongoing or future interaction with the ratees. Instead, agreeable raters are likely to be more lenient in their ratings as a function of their disposition to be kind to others. Kane et al. (1995) also suggested that because conscientious individuals tend to be ethical, responsible,
deliberate, and organized, conscientious raters are likely to be less prone to leniency bias and take principled approach to conduct performance appraisal in an objective manner.

Previous study results have generally corroborated Kane et al.’s (1995) propositions. For example, Bernardin and colleagues (e.g., Bernardin, Cooke, & Villanova, 2000; Bernardin, Tyler, & Villanova, 2009) found that rater agreeableness was associated with rating inflation and rating inaccuracy, and that rater conscientiousness was associated with lower level and more accurate ratings. Moreover, Bernardin and colleagues also found that raters with a combination of high agreeableness and low conscientiousness provided the most lenient and least accurate ratings compared to raters who were just high on agreeableness or low on conscientiousness. Research has also shown that conscientious raters are less likely to be influenced by the context of performance appraisal. In a study by Roch et al. (2005), the authors asked college students to evaluate their instructor’s teaching performance and manipulated whether their ratings could be identified or not (name vs. no name on the evaluation sheet). The results showed that there was no significant difference in the level of ratings between high and low conscientious students in the anonymous condition, but for students who were led to believe that they could be identified, low conscientious students were significantly more lenient, whereas high conscientious students did not show significant difference than high conscientious raters in the anonymous condition. Taken together, research has generally shown that accountability context and individual differences in rater personality both play an important role in determining the ratings that raters provide.

With respect to the goal-directed performance appraisal perspective (Murphy & Cleveland, 1991, 1995), the general framework of the relationship between
accountability context, rater personality, and performance ratings can be described as a mediated model (see Figure 1; Murphy, 2008). That is, contextual demands in different accountability situations are likely to motivate raters to approach performance appraisal with rating goals that satisfy those demands. Similarly, individual differences in rater personality are likely to motivate raters to approach performance appraisal with rating goals that align with their dispositional tendency in interacting with others.
Figure 1. Mediation model of the performance – performance rating relationship. Contextual = contextual factors; Individual = rater individual difference characteristics; Error = measurement error.
However, theoretical account of the effect of person-situation interaction on organizational behavior (trait activation theory; Tett & Burnett, 2003) suggests that the combined effect of rating context and rater characteristics on performance ratings may be better described as a moderated-mediation model. That is, the effect of rater characteristics on how raters choose to pursue performance appraisal is likely to be moderated by the context in which performance appraisal takes place.

In the following section, I will describe Tett and Burnett’s (2003) trait activation theory and how it relates to my study hypotheses on the effect of accountability context and rater personality on performance ratings.

**Person-Situation Interaction in Performance Appraisal**

Tett and Burnett’s (2003) trait activation theory posits that trait-relevant situational cues may moderate the strength of the relationship between individuals’ personality and behavior. That is, situations may provide a context for personality-expressive behavior, in which case the cued personality becomes more predictive of behavior in that situation. Then, it is reasonable to hypothesize that if accountability context demands rating behavior that aligns with raters’ dispositional tendencies, accountability context would cue raters to display more personality-expressive rating behavior. Thus, when there is a match between rating behaviors that are invoked by the accountability context and rating behaviors that are invoked by the individual differences in rater personality, rater personality should be more predictive of rating behaviors in the presence of the accountability context.

In the current study, I examined how the match between rating goals that are invoked by the accountability context, and rating goals that are invoked by the individual differences in rater personality moderates the mediating effect of rater intentions on the relationship between rater personality and performance ratings.
Specifically, based on the principles of the trait activation theory (Tett & Burnett, 2003), I expect that accountability context would influence performance ratings by moderating the relationship between rater personality and rating goals that raters choose to pursue in performance appraisal.
Chapter 3

Study Overview

There were mainly two parts to the current study. The first part of the study focused on testing the theoretical models discussed above. Specifically, based on the goal-directed performance appraisal perspective, I examined whether accountability context and rater personality affects performance ratings (level of ratings and rating accuracy) through the influence that these two factors are expected to have on raters’ rating intentions. Then, I tested the theoretical model proposed in the current study that accountability context would moderate the mediating effect of rater intentions on the relationship between rater personality and performance ratings. In the second part of the study, I tested the effect that accountability context and rater personality had on the mean level and accuracy of the performance ratings.

Typology of Rating Goals

Previous efforts to categorize rating goals have emphasized qualitative approaches to assembling raters’ rating intentions (e.g., interview; Kerst, 1993; Longenecker et al., 1987), but subsequently grouping them through subjective judgment of substantive commonalities among the different rating intentions that are found (e.g., Murphy & Cleveland, 1991, 1995; Spence & Keeping, 2011, 2013). For example, based on a comprehensive review of the literature on rating distortion behaviors, Spence and Keeping (2011) suggested that rating motivations that are discussed in the performance appraisal literature may be classified into two broad categories: (1) attempting to attain a valued end (motivation to enhance one’s image, motivating ratees, procuring desired results/resources; e.g., Cleveland & Murphy, 1992; Longenecker et al., 1987; Tziner, Latham, Price, & Haccoun, 1996), or (2) attempting to avoid an undesired end (motivation to avoid confrontations with other
employees, motivation to avoid damaging rater-ratee relationship; e.g., Cleveland & Murphy, 1992; Harris, 1994; Longenecker et al., 1987; Wong & Kwong, 2007).

Longenecker et al. (1987) found diverse, but substantively common managerial accounts of political considerations that motivate managers to inflate or deflate performance ratings. Specifically, Longenecker et al. found that managers often inflated performance ratings to effectively maintain or increase the ratee’s level of job performance, to maximize the rewards that the ratee would be eligible to receive, avoid publicly displaying low performance ratings that may be viewed by people outside of the organization, avoid confrontation with the ratees, or to reward ratees who have shown improvement in performance. Longenecker et al. also found that managers often engaged in deliberate rating deflation to shock underperforming employees to perform better, to send a message that the employee should consider leaving the job (or the organization), and to build a documented track record of poor-performing employees’ job performance.

In a review by Spence and Keeping (2013), the authors suggested that four common rating intentions discussed in the performance appraisal literature were accuracy (i.e., rating objectively), avoiding conflict (i.e., appeasing the ratee), benevolence (i.e., providing helpful, considerate ratings), and impression management (i.e., self-enhancement).

**Typology of rating goals in the current study.** I carefully examined the models of intentional rating behavior mentioned above (as well as the sources of these models; Harris, 1994; Murphy & Cleveland, 1995; Spence & Keeping, 2011, 2013), and decided to orchestrate the current study around the prediction of four rating goals: (1) conflict-avoidance goal, (2) critical goal, (3) equity fairness goal, and (4) interpersonal goal. I decided to focus on these four particular goals because they were
commonly discussed in the literature on intentional rating behavior, and they were theoretically relevant to the behavioral tendencies associated with the rater personality traits examined in the current study (agreeableness, conscientiousness, and neuroticism). Of course, this list of rating goals is not meant to be comprehensive nor does it represent the full range of rating goals that are related to the dispositional tendencies associated with the rater personality traits examined in the current study. In the sections that follow, I will explain the four rating goals in more detail, and in the hypothesis section, explain why I expect them to be related to the personality traits examined in the current study.

**Conflict-avoidance goal.** As mentioned above, one of the most commonly discussed rating intentions discussed in the literature on intentional rating behavior is motivation to avoid conflict with the ratees (e.g., Harris, 1994; Murphy & Cleveland, 1991, 1995; Spence & Keeping, 2013). For example, Kane and Kane (1989) surveyed senior personnel executives in the *Fortune* 500 firms and found that intentional rating inflation resulting from raters wanting to avoid confrontations with ratees was the most widely mentioned problem in performance appraisals. Similarly, in Bretz et al.’s (1992) review of surveys on performance appraisal practices, as well as anecdotal accounts regarding performance appraisal practice in military, public, and private sector organizations, rater intention to appease the employee was consistently mentioned. In addition to motivation to avoid conflict with ratees, research has also shown that raters are concerned about appealing to the preferences of an audience of higher status when their rating behaviors are monitored closely, or if raters perceive that their rating behaviors can be identified (e.g., concern about appearing too lenient; Murphy & Cleveland, 1995). Based on these findings, in the
current study, conflict-avoidance goal represents rater intention to provide performance ratings that appeal to the demands of the audience.

**Critical goal.** Rater intention to critically evaluate employee performance was also commonly reported in qualitative data on intentional rating behavior (e.g., Gioia & Longenecker, 1994; Kerst, 1993; Longenecker et al., 1987). For example, raters sometimes deliberately provided negative ratings to jolt an employee to perform better, to teach a poor performing employee a lesson, or send an indirect message that the employee is not fit for the job (Longenecker et al., 1987). According to Longenecker et al. (1987), this tactic is expected to become more common because organizations must be able to provide clear documentation of employees’ track record of poor performance in litigation cases associated with unlawful termination, which are becoming increasingly prevalent.

**Equity fairness goal.** There are broadly two standards to represent fairness in performance appraisal: (1) equality principle (Deutsch, 1985), and (2) equity principle (Adams, 1965). Equality principle refers to rewards being distributed equally among group members, which in the context of performance appraisal, would be associated with egalitarian, less discriminatory rating patterns (i.e., everyone receiving the same rating); equity principle refers to rewards being distributed according to a member’s contribution to the group, which in the context of performance appraisal, would be associated with a discriminatory rating pattern in which employees receive a rating that represents their task input. In the current study, fairness goal was represented as rater motivation to follow the equity principle.

Providing accurate performance ratings that adhere to objective performance standards, and reflect ratees’ true level of job performance is the supposed purpose of performance appraisal. Although the nature of an organizational environment makes
it a difficult rating intention to pursue, supervisors often reported that providing accurate ratings is an important goal that they engage in when conducting performance appraisal (Kerst, 1993).

**Interpersonal goal.** Interpersonal goal represents rater intention to approach performance appraisal benevolently (e.g., providing ratings that are considerate of ratees’ reaction to the ratings, being considerate of how ratings might influence ratees’ morale). Qualitative data clearly indicates that raters are often concerned about how ratees will react to their ratings (e.g., concerns about decreasing employee morale, concerns about demotivation; Bjerke et al., 1987; Longenecker et al., 1987).

Theoretical models of intentional rating behavior proposed that such benevolent approach to evaluating employees should explain why rating leniency is so prevalent in practice (Harris, 1994; Murphy & Cleveland, 1995; Spence & Keeping, 2013).

**Study Hypotheses**

**Effect of Rater Personality on Rating Goals and Performance Ratings**

**Agreeableness.** Agreeable individuals are described as being kind, generous, cooperative, and concerned about the well-being of others (Costa & McCrae, 1992). In relation to these behavioral tendencies, agreeable individuals are also motivated to receive social approval (Meier, Robinson, & Wilkowski, 2006), and avoid interpersonal conflict (Jensen-Campbell, Knack, Waldrip, & Campbell, 2007). If rater personality affects performance ratings through affecting how raters choose to pursue performance appraisal, agreeable raters would be more likely to pursue rating goals that closely align with their tendency to be kind to others and to avoid social conflict. Additionally, because these rating goals describe motivation to be generous, rater agreeableness is likely to be predictive of higher level and less accurate performance ratings.
Hypothesis 1: Rater agreeableness will predict performance ratings (higher rating and lower rating accuracy) through motivation to pursue interpersonal goal and conflict-avoidance goal.

Neuroticism. Neurotic individuals are generally described as being anxious, worried, insecure, and susceptible to stress (Costa & McCrae, 1992). Related to these behavioral tendencies, neurotic individuals also tend to react more negatively to stressful situations (Farber & Spence, 1956; Spielberger, Gorsuch, & Lushene, 1970). Given these attributes that are associated with neuroticism, in performance appraisal situations, neurotic raters are more likely to be concerned about their own security rather than fulfilling their role as a fair rater. Then, it is reasonable to suggest that neurotic raters would be motivated to provide ratings that help them evade potential confrontation with others about the ratings that they provide. If this is true, the effect of rater neuroticism is likely to vary depending on the demands of the appraisal situation, which in the context of accountability, should vary depending on who the audience is. That is, neurotic raters might be willing to significantly inflate or deflate the ratings that they provide in order to avoid facing conflict with the audience to whom they are held accountable.

Hypothesis 2a: Rater neuroticism will predict performance ratings through motivation to pursue conflict-avoidance goal.

However, it is also reasonable to suggest that neurotic raters might actively seek opportunities to be critical of others. Based on consistent evidence of a strong correlation between neuroticism and negative affectivity, researchers suggested that these two factors capture a common underlying construct (Eysenck, 1970; Gotlib, 1984; Watson & Clark, 1984; Zurawski & Smith, 1987). Negative affectivity may be described as a general factor of subjective distress that subsumes a wide range of
negative mood states (e.g., anxiety, hostility) and depressive mood (e.g., sadness, loneliness; Watson, Clark, & Carey, 1988), which has also been shown to be associated with negative affective states, such as anger, scorn, and revulsion. Then, in the context of performance appraisal, neurotic raters might be more likely to be motivated to scorn the ratees about their ineffective performance episodes and be critical of their performance in general, which should be predictive of deflated and less accurate performance ratings.

*Hypothesis 2b: Rater neuroticism will predict performance rating (lower rating and lower rating accuracy) through motivation to pursue critical goal.*

**Conscientiousness.** A principled approach to performance appraisal includes providing performance ratings that accurately and objectively reflects ratees’ level of job performance. Conscientious individuals are generally described as ethical, responsible, persistent, deliberate, and organized (Costa & McCrae, 1992). Given these attributes that are associated with conscientiousness, it is reasonable to suggest that conscientious raters are likely to approach performance appraisal with rating goals that align with their dispositional tendency to approach tasks in a responsible manner, even if it involves facing difficult circumstances, such as providing ratings that ratees may not appreciate. Accordingly, rater conscientiousness is likely to be predictive of performance ratings through engagement in equity fairness goal and critical goal in performance appraisal process. Because these rating goals describe motivation to be objective and critical (if necessary), rater conscientiousness is likely to be predictive of lower level of rating, and more accurate rating.

*Hypothesis 3: Rater conscientiousness will predict performance ratings (lower rating level and higher rating accuracy) through motivation to pursue equity fairness goal and critical goal.*
Effect of Accountability Context on Performance Ratings

In performance appraisal, there are typically two groups of audiences to whom raters are held accountable: (1) ratees who receive the ratings, and (2) supervisors who oversee the performance appraisal process and approve the ratings (Bernardin & Villanova, 1986; Klimoski & Inks, 1990; Mero et al., 2007). Ratees and supervisors occupy different positions in the performance appraisal process, and each party presents a different motivational context to raters. Specifically, ratees typically represent a pull towards inflated, inaccurate ratings, whereas supervisors typically represent a pull towards lower, accurate ratings (with the assumption that supervisors expect accurate ratings to help operational decisions made based on the ratings; Roch & McNall, 2007).

Accountability to ratees. When raters must justify their ratings to the ratee, a clear coping strategy is to provide positive ratings because they are much more likely to please the ratees than negative ratings. In fact, results from several performance appraisal studies have shown that rating inflation occurs, at least in part, as a function of rater motivation to keep ratees satisfied. For example, it has been shown that raters inflate ratings out of politeness towards the ratees (Brown & Levinson, 1987), empathy towards the negative consequences that may befall on the ratees with low ratings (Schlenker & Britt, 2001), and to protect ratee’s feelings (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996). Studies have also suggested that rating inflation may be attributed to rater motivation to avoid conflict with ratees. That is, the most likely instance in which ratees are expected to confront the raters is when they receive a low rating, so raters inflate ratees’ scores to avoid getting involved in a complicated argument with the ratees (Fisher, 1979; Longenecker et al., 1987; Waung & Highhouse, 1997).
Based on these findings, it is reasonable to suggest that when raters are held accountable to ratees, raters are more likely to provide inflated, inaccurate ratings because they are more likely to be deliberately generous (i.e., pursue interpersonal goal), and be motivated to avoid upsetting the ratees (i.e., pursue conflict-avoidance goal) compared to a neutral context (baseline condition) in which there is less contextual pressure to satisfy the ratees.

Hypothesis 4a: In the ratee accountability condition, raters are likely to engage more in interpersonal goal and conflict-avoidance goal than in the baseline condition; this is expected to result in higher, inaccurate ratings compared to ratings in the baseline condition.

Accountability to supervisors. When raters must justify their ratings to an audience of higher authority, and raters are not provided with information about what ratings the audience would prefer, raters are likely to account for their ratings by providing ratings that accurately represent the ratees’ level of job performance. That is, without the knowledge about the preference of the audience, the most defendable ratings that place the least amount of liability on the raters are accurate ratings (Tetlock, 1985). Additionally, higher authority of the audience should also motivate raters to want to subsume to their demands. Thus, I expect that raters held accountable to supervisors would be more likely to pursue equity fairness goal and critical goal compared to raters in a neutral context (i.e., baseline condition) where there is less contextual pressure to satisfy the supervisors. These rating goals should also result in less inflated and more accurate ratings.

Hypothesis 4b: In the supervisor accountability condition, raters are more likely to be motivated to pursue equity fairness goal and critical goal than in
the baseline condition; this is expected to result in lower, accurate ratings compared to ratings in the baseline condition.

Moderating Effect of Accountability on the Mediating Effect of Rating Goal

I discussed earlier that accountability context may moderate the relationship between rater personality and rating goal such that accountability context may strengthen the relationship between rater personality and rating goals to the extent that accountability context prompts raters to elicit their dispositional tendencies. Then, if being held accountable to ratees motivates raters to approach performance appraisal with empathy towards the ratees (interpersonal goal) and to avoid conflict with them (conflict-avoidance goal), and if rater agreeableness and neuroticism are also predictive of engagement in interpersonal goal and conflict-avoidance goal in performance appraisal, it is reasonable to suggest that rater agreeableness and neuroticism would be more predictive of engagement in interpersonal goal and conflict-avoidance goal when raters are held accountable to ratees for their ratings compared to when raters are not held accountable. Similarly, if being held accountable to supervisors motivates raters to be objective (equity fairness goal) and critically assess ratees (critical goal), and if rater conscientiousness is predictive of engagement in equity fairness goal and critical goal in performance appraisal, it is reasonable to suggest that rater conscientiousness would be more predictive of engagement in equity fairness goal and critical goal when raters are held accountable to supervisors for their ratings compared to when raters are not held accountable.

Hypothesis 5: Predictability of rating goals by rater personality will be moderated by accountability context such that predictability will be strengthened when rater personality and accountability context elicits the same rating goals.
Rater Personality × Accountability Context Effect on Performance Ratings

Agreeableness × Accountability interaction. If agreeable raters are more likely to provide inflated and inaccurate ratings resulting from their inclination to engage in interpersonal goal and conflict-avoidance goal, and if being held accountable to ratees also exerts situational pressure on raters to engage in interpersonal goal and conflict-avoidance goal, it is reasonable to suggest that rater agreeableness and accountability context may have a multiplicative effect on overall performance rating and rating accuracy. Specifically, higher rater agreeableness in a situation where raters are held accountable to ratees is likely to lead to even more rating inflation and inaccurate ratings than in a neutral situation where there is unlikely to be much situational pressure on the raters to elicit their agreeable tendencies.

Hypothesis 6a: Rater agreeableness and accountability context will have multiplicative effect on overall performance rating and rating accuracy such that additional unit of rater agreeableness will result in higher rating and lower rating accuracy in the ratee accountability condition than in the baseline condition.

Neuroticism × Accountability interaction. I hypothesized that accountability context would moderate how neurotic raters would approach performance appraisal. Because it is reasonable to expect that higher ratings would be easier to justify to ratees, neurotic raters are more likely to provide inflated and inaccurate ratings when they are held accountable to ratees than when there is no audience for their ratings. Similarly, because raters are expected to provide ratings that are objective when they are held accountable to an authority without an explicit preference, neurotic raters are more likely to provide lower and more accurate ratings
when they are held accountable to supervisors for their ratings than when there is not

Hypothesis 6b: Raters high in neuroticism are likely to provide higher and
inaccurate ratings in the ratee accountability condition than in the baseline
condition; raters high in neuroticism are also likely to provide lower and
accurate ratings in the supervisor accountability condition than in the
baseline condition.

Conscientiousness × Accountability interaction. If conscientious raters are
more likely to provide less inflated and more accurate ratings because they are more
likely to be motivated to provide objective ratings, and if being held accountable to
authority elicits motivation to provide objective ratings, it is reasonable to suggest that
there could be a multiplicative effect between rater conscientiousness and
accountability context on performance ratings. Specifically, higher rater
conscientiousness in the supervisor accountability condition is likely to lead to even
more objective and accurate ratings compared to higher rater conscientiousness in a
situation where there is no situational pressure on raters to elicit conscientious rating
tendencies.

Hypothesis 6c: Rater conscientiousness and accountability context will have
multiplicative effect on overall performance rating and rating accuracy such
that additional unit of rater conscientiousness will result in lower rating and
higher rating accuracy in the supervisor accountability condition than in the
baseline condition.
Chapter 5
Method

Sample

The sample for this study was limited to organizational supervisors with managerial experience. Participants were recruited through Qualtrics Panel, a service from Qualtrics that allows researchers to request a specific sample for their research (please visit www.qualtrics.com for more information). Participants were rewarded for their participation in the form of online points, which they can redeem in various forms of rewards (e.g., airline mileage, online store credit).

In requesting a sample for the current study, I asked for a 50/50 gender ratio for the study sample (see Table 1 for gender ratio of the study sample) when it seems that there is a greater proportion of male supervisors than female supervisors in the workplace. However, previous studies that have looked at the effect of accountability (e.g., Mero et al., 2007), rater personality (e.g., Bernardin et al., 2009), and rating goals (e.g., Wang, Wong, & Kwong, 2010) did not find significant gender effect on the study results and had study samples that were evenly balanced in gender. Additionally, a number of studies found that the interaction between a rater and a ratee’s gender (i.e., match/mismatch in rater – ratee gender) accounted for an extremely small amount of variance in ratings (e.g., Pulakos, White, Oppler, & Borman, 1989; Shore & Thornton, 1986; Thompson & Thompson, 1985). The correlations between gender and the main study variables were non-significant in the data for the current study, suggesting that gender did not meaningfully influence the study results.

A total of 188 participants provided complete data for analysis. Participants ranged in age from 18 to 76 years old. The participants generally had an extensive
supervisory experience (mean number of years in a supervisory position = 10.8, \( SD = 9.2 \)), and experience in managing employees in an organization (mean number of employees currently managing [or most recently managed] = 25.1, \( SD = 60.0 \), median = 10). The distribution of the number of employees that the participants are currently (or most recently) managing is positively skewed (skewness = 5.89) with a minimum of 7 employees to a maximum of 500 employees. More detailed demographic information on gender and ethnicity of the participants is shown in Table 1. The demographics of the current sample may not be representative of the managerial population demographics, but this should not seriously limit the ability of the current study’s attempt to identify the general trends and principles of the effect that accountability and rater personality has on raters’ intentional performance appraisal behavior.
<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>% Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>96</td>
<td>51.1</td>
</tr>
<tr>
<td>Male</td>
<td>92</td>
<td>48.9</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African-American</td>
<td>21</td>
<td>11.2</td>
</tr>
<tr>
<td>Asian</td>
<td>12</td>
<td>6.4</td>
</tr>
<tr>
<td>Caucasian</td>
<td>131</td>
<td>69.7</td>
</tr>
<tr>
<td>Hispanic</td>
<td>14</td>
<td>7.4</td>
</tr>
<tr>
<td>Native American</td>
<td>14</td>
<td>7.4</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>4.2</td>
</tr>
</tbody>
</table>
Procedure and Experimental Manipulation

Participants signed up for a study entitled “Methods of Performance Appraisal.” Participants were told that the purpose of the study was to investigate the different approaches that supervisors take in evaluating employee performance. All participants were told at the beginning of the study that they would watch a 10-minute performance episode of a male regional manager of a sales company dealing with a situation brought forth by his direct report, and that they would be asked to provide an overall evaluation of the manager’s performance in the performance episode.

Participants were randomly assigned into one of three study conditions: ratee accountability condition \((N = 64)\), supervisor accountability condition \((N = 59)\), and baseline condition \((N = 65)\). To induce the accountability manipulation, participants in the ratee accountability condition were told that they would be asked to write a brief statement to the manager they evaluated in the experiment justifying their ratings. Similarly, participants in the supervisor accountability condition were told that they would be asked to write a brief statement to the researchers conducting the study justifying their ratings. Participants in these two accountability conditions were also told that the accountable group (ratee in the ratee accountability condition, researchers in the supervisor accountability condition) would judge the justification statements, and reward a participant whose justification is judged to be the most effective with a $100 gift certificate to an online retail store. Participants in the baseline condition were not given any instructions regarding providing justification for their ratings, and were told that one participant would be randomly selected for a $100 gift certificate to an online retail store.

After participants finished watching the performance episode video and completed the performance evaluation, I asked them to rate the extent to which they
approached the evaluation with each of the four performance rating goals mentioned above (see Appendix). Participants then completed a self-reported measure of the Big Five, and perceived accountability items. Afterwards, participants were debriefed about the actual purpose of the study and that they would not actually have to provide any justification for their ratings. Participants were also thanked for their participation.

**Materials**

**Performance episode.** A 10-minute video of a performance episode from an executive developmental assessment center was used as the performance stimulus in the current study (this performance episode has been used in previous studies; Gorman & Rentsch, 2009; Hoffman et al., 2012). The performance episode depicted a role-play exercise, in which a candidate participating in the assessment center took on the role of a manager and was asked to interact with a subordinate played by a trained assessor. Participants in the current study were asked to evaluate the manager’s performance in the episode (the manager was the only person being shown in the video).

**Measures**

**Big Five personality traits.** 12 items from the Mini-IPIP (Donnellan, Oswald, Baird, & Lucas, 2006) were used to assess participants’ agreeableness, conscientiousness, and neuroticism (4 items per factor). Items were on a 5-point Likert-type scale ranging from 1 (very inaccurate) to 5 (very accurate). Alpha coefficients for scores on each personality dimension were .67 for agreeableness, .70 for conscientiousness, and .62 for neuroticism.

**Rating goals.** Murphy et al. (2004) developed their rating goal questionnaire by conceptually grouping the rating goals found in Kerst’s (1993) study into four
factors (identifying weaknesses, identifying strengths, providing fair/accurate ratings, and motivating ratees), and quantitatively testing the model through factor analysis. Similar to Murphy et al.’s (2004) approach, I adapted statements from qualitative data on intentional rating behavior (e.g., Longenecker et al., 1987) and items from previous questionnaires of political motivations in performance appraisal (e.g., Tziner et al., 1996) to develop a 20-item questionnaire that purport to assess the four rating goals examined in the current study (5 items per goal). Participants were asked to indicate the extent to which they engaged in each intentional rating behavior statement when they were evaluating the manager’s job performance in the experiment. Items were on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). Sample items include “I was motivated to be kind to the manager receiving my rating,” (interpersonal goal) and “I was motivated to rate the manager’s performance in a manner that reflected his performance” (equity fairness goal). Alpha coefficients were .92 for critical goal, .94 for conflict-avoidance goal, .91 for equity fairness goal, and .88 for interpersonal goal. Confirmatory factor analytic (CFA) test of the proposed four-factor structure showed good model fit to the data (CFI = .91; RMSEA = .09).

**Rating form.** Participants were asked to rate the ratee’s overall job performance on a 7-point Likert-type scale ranging from 1 (very ineffective) to 7 (very effective).

**Perceived accountability.** After the experiment, participants were asked to indicate their feelings of accountability on two items. The items asked, “I felt responsible for my ratings,” and “I felt answerable for my ratings.” Items for participants in the ratee accountability condition ended with a statement “to the manager in the video,” and items for participants in the supervisor accountability
condition ended with a statement “to the researchers.” Items were on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree).

**Dependent Variables**

**Rating accuracy.** Rating accuracy was operationalized using Cronbach’s (1995) $D^2$ index. Put simply, $D^2$ index defines accuracy as the squared distance between the observed criterion score and the true criterion score, where true criterion score is operationalized by criterion score from subject matter experts (SMEs). The true score in the current study was collected from three SMEs using the procedures recommended by Sulsky and Balzer (1988). The SMEs were upper-level graduate students in I/O psychology, each of who had been trained as an assessment center assessor (intensive 30-hour training over 6 days, followed by an annual day-long review training). The SMEs had an average of 3 years of assessment center experience, during which they regularly conducted assessments for both administrative and developmental purposes (see Gorman & Rentsch, 2009 and Hoffman et al., 2012 for more information about the SMEs). After each SME independently rated the performance episode presented in the video, they met to achieve a consensus rating. This consensus rating was used as the true score. The true overall job performance rating in the experiment was originally 2.75 out of 5.0. However, it was adjusted to 3.78 as I adjusted the overall performance rating into a 7-point scale.

**Analysis**

**Perceived accountability.** I examined the difference in accountability perception item scores among the experimental conditions. Effective accountability manipulation should result in higher accountability score for participants in the accountability condition groups than for participants in the baseline condition. I also
examined how directly measured perceived accountability affected the study results and discussed their implications.

**Mediation and moderated-mediation model.** I tested the hypothesized mediation and moderated-mediation effects using the SPSS macros developed by Hayes, Preacher, and colleagues (Hayes & Preacher, 2014; Preacher & Hayes, 2008; Preacher, Rucker, & Hayes, 2007). In recent methods of mediation analysis, a mediation effect is quantified as a single number for which confidence intervals and significance tests can be calculated (MacKinnon, Coxe, & Baraldi, 2012).

Specifically, the effect of $X$ (predictor) to $Y$ (dependent variable) via $M$ (mediator) is quantified as the product of the regression coefficient between $X$ and $M$ (i.e., path $a$; Baron & Kenny, 1986) and the regression coefficient between $M$ and $Y$ (i.e., path $b$; Baron & Kenny, 1986). Dividing $ab$ by its standard error and comparing this value to a standard normal ($z$) distribution calculates the statistical significance of $ab$, for which the null hypothesis is $H_0: ab = 0$.

The most commonly used standard error to calculate the statistical significance of a mediation effect is Sobel $\sigma$ (significance test of a mediation effect using Sobel $\sigma$ is called Sobel test; Sobel, 1982). However, a problem with the Sobel test is that it assumes normal distribution of the mediation effect, which is only true under the circumstances that both $a$ and $b$ are equal to zero (MacKinnon, Lockwood, & Williams, 2004). Thus, Sobel test is often conducted under violation of the assumed normality of the mediation effect, which results in reduced power to detect a significant mediation and imbalanced confidence limits (MacKinnon et al., 2004; MacKinnon et al., 2012).

To address this issue with the Sobel test, researchers recommend testing for significance of mediation effects using bootstrap method (Cheung & Lau, 2008;
MacKinnon et al., 2012). Bootstrapping method involves taking a sample size of $N$ cases (where $N$ is the size of the original sample; Edwards & Lambert, 2007) from the original sample with replacement, so a given case may be selected as a part of a bootstrapped sample multiple times. Then, all estimates in the mediation model can be re-estimated using a new sample of $N$. This re-sampling and re-estimation process is repeated $k$ times, where $k$ is recommended to be at least 1,000 (Preacher & Hayes, 2008). The distribution of these $k$ estimates then serves as nonparametric approximation of the sampling distribution for each of the model estimates. Then, sorting the $k$ values for each model estimates, the $(.5\alpha)k^{th}$ and $1 + 1(-.5\alpha)k^{th}$ values, where $\alpha$ is the desired Type I error rate, marks the lower and upper bounds of a 100 $(1 – \alpha)$% confidence interval for the model estimates (Preacher & Hayes, 2008). An effect is considered statistically significant if zero is not included in the confidence interval (MacKinnon et al., 2004). In addition to addressing the power and imbalanced confidence interval issues in Sobel test, bootstrap method is useful because it provides variability in effect size for each model estimates, which allows better understanding of the strength of the modeled estimates than just a binary significance decision regarding the estimate (Harlow, Mulaik, & Steiger, 1997).

Among the numerous bootstrap methods that are available (e.g., percentile method, bootstrap-$t$ method), I used the bias-corrected (BC) bootstrap method to estimate the confidence interval as it has been shown to produce the most accurate confidence limits, most accurate Type I error, and provides the strongest power for detecting a significant mediation effect (Cheung & Lau, 2008). In addition to testing the significance of the mediation effect using bootstrap method, statistical significance of specific paths in the mediation model ($a, b, c, c'$) was tested using Sobel test.
The general framework for moderated mediation builds on the basic mediation model by adding product terms involving a moderator variable \( Z \) that shows how \( Z \) influences the paths between \( X \), \( M \), and \( Y \), and any combination of these paths in the mediation model (i.e., indirect effects and total effects). Hypotheses in the current study were centered on *first stage moderation*, which applies the moderating effect of \( Z \) on the first stage of the indirect effect of \( X \) on \( Y \) (a path). In the context of the current study, first stage moderation tests the moderating effect of \( Z \) (accountability context) on the effect of \( X \) (rater personality) on \( M \) (rating goal).

**Mediation analysis with a multicategorical independent variable.** However, the process for testing a mediation effect with a multicategorical independent variable is a bit different. With a multicategorical independent variable, there can be no single regression slope value that represents the effect of an independent variable to a mediator or a criterion. Instead, to fully represent the effect of a categorical variable with \( k \) mutually exclusive categories, \( k - 1 \) parameter estimates are needed (Hayes & Preacher, 2014).

In a tutorial paper by Hayes and Preacher (2014), the authors demonstrated a method of conducting a mediator analysis with a multicategorical independent variable. In this method, a category in the multicategorical independent variable is selected as a reference category, and parameter estimate that represents the difference in the effect of the comparison groups relative to the effect for the reference group are calculated. I tested the effect of accountability context on performance ratings (Hypothesis 4) using Hayes and Preacher’s (2014) method. I chose the baseline condition as the reference group, and the two accountability conditions as the comparison groups. This analysis was also tested using the BC bootstrap CI.
Mean comparisons in performance ratings. I classified the raters into being high vs. low for each personality trait variable through median split on the participants’ scores on the Mini-IPIP. I then conducted a series of between-group comparisons to examine the relative difference in the mean overall performance rating and rating accuracy between different combinations of the level of rater personality, accountability condition, and their interaction.
CHAPTER 5

Results

Perceived Accountability

Because I expected participants in the two accountability conditions to report being more accountable for their ratings than participants in the baseline condition, I conducted a between-subjects ANOVA with an a priori contrast between the mean accountability score for participants in the two accountability conditions (ratee condition and supervisor condition), and the mean accountability score for participants in the baseline condition. The result, however, showed a non-significant difference, $t(187) = 1.60, p = .11$. The overall ANOVA was also non-significant. In fact, perceived accountability was high for all of the study conditions ($M = 4.26, SD = .78$ for ratee accountability condition; $M = 4.07, SD = .87$ for supervisor accountability condition; $M = 4.35, SD = .69$ for baseline condition; mean scores were converted to a 5-point scale).

Mediation and Moderated Mediation Analysis

Table 2 shows the means, standard deviations, and zero-order correlations for the main study variables. Both conflict-avoidance goal and interpersonal goal was associated with higher rating and lower rating accuracy (positive correlation indicates lower accuracy because higher Cronbach’s $D^2$ indicates lower accuracy). Surprisingly, equity fairness rating goal was associated with lower rating accuracy, whereas critical rating goal was not significantly correlated with overall rating or rating accuracy. Perceived accountability was significantly associated with higher level of socially desirable personality traits (agreeableness [$r = .21$], conscientiousness [$r = .31$]), and significantly associated with lower level of neuroticism ($r = -.29$), a socially undesirable trait. Similarly, accountability scores
were significantly associated with rating intentions that are associated with responsible performance appraisal (interpersonal goal \( r = .16 \), critical goal \( r = .27 \), fairness goal \( r = .30 \)). Accountability scores were not correlated with age, number of years full-time, number of years in supervisor position, and number of employees supervised. Overall performance rating was significantly correlated with rating accuracy, suggesting that higher rating was associated with lower rating accuracy.

Additionally, overall performance rating and rating accuracy were regressed on rater personality traits and rating goals. Results showed that rater personality traits did not significantly predict overall performance rating \((\Delta R^2 = .02, p = .35)\) or rating accuracy \((\Delta R^2 = .01, p = .54)\), whereas rating goals produced significant change in \( R^2 \) for both overall performance rating \((\Delta R^2 = .31, p < .001)\) and rating accuracy \((\Delta R^2 = .26, p < .001)\).
Table 2
Means, Standard Deviations, and Zero-Order Correlations between the Main Study Variables

<table>
<thead>
<tr>
<th></th>
<th>M (SD)</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
<th>4.</th>
<th>5.</th>
<th>6.</th>
<th>7.</th>
<th>8.</th>
<th>9.</th>
<th>10.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Agreeableness</td>
<td>3.88 (.65)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2. Conscientiousness</td>
<td>4.06 (.67)</td>
<td>.50**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3. Neuroticism</td>
<td>2.43 (.73)</td>
<td>-.41**</td>
<td>-.50**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4. Interpersonal Goal</td>
<td>3.52 (.79)</td>
<td>-.01</td>
<td>-.10</td>
<td>.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5. C-A Goal</td>
<td>2.61 (1.11)</td>
<td>-.24**</td>
<td>-.28**</td>
<td>.25**</td>
<td>.48**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6. Critical Goal</td>
<td>3.75 (.84)</td>
<td>.04</td>
<td>.13</td>
<td>-.13</td>
<td>.03</td>
<td>-.15*</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7. Fairness Goal</td>
<td>4.13 (.72)</td>
<td>.33**</td>
<td>.36**</td>
<td>-.29**</td>
<td>.05</td>
<td>-.35**</td>
<td>.45**</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8. Accountability</td>
<td>4.23 (.78)</td>
<td>.21**</td>
<td>.31**</td>
<td>-.29**</td>
<td>.16*</td>
<td>-.03</td>
<td>.27**</td>
<td>.30**</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9. Accuracy</td>
<td>2.69 (2.78)</td>
<td>.08</td>
<td>.02</td>
<td>.04</td>
<td>.35**</td>
<td>.22**</td>
<td>-.04</td>
<td>.28**</td>
<td>.17*</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10. Performance</td>
<td>4.68 (1.37)</td>
<td>-.05</td>
<td>-.06</td>
<td>.13</td>
<td>.53**</td>
<td>.28**</td>
<td>-.13</td>
<td>.06</td>
<td>.06</td>
<td>.57**</td>
<td>-</td>
</tr>
</tbody>
</table>

Note. N = 188. M = mean; SD = standard deviation; C-A goal = conflict-avoidance goal. ** correlation significant at p < .01. * correlation significant at p < .05.
Agreeableness. Table 3a shows the mediation analysis results for rater agreeableness. Results for the $a$ paths of the proposed mediation model showed that rater agreeableness significantly predicted engagement in conflict-avoidance goal, but agreeableness was actually predictive of lower engagement in conflict-avoidance goal. Also contrary to my expectations, rater agreeableness did not significantly predict engagement in interpersonal goal. Results for the $b$ paths of the model showed that interpersonal goal was significantly predictive of higher overall rating and lower rating accuracy (as expected), but conflict-avoidance goal did not significantly predict either overall performance rating or rating accuracy. Table 3a also shows the estimates and 95% confidence intervals for the bootstrapped total and specific indirect effects. Contrary to my expectations, the hypothesized indirect effects were non-significant for both overall performance rating and rating accuracy.
Table 3a

Simple Mediation Analysis Results for Rater Agreeableness

<table>
<thead>
<tr>
<th>Variable</th>
<th>b</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a path</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness → Interpersonal Goal</td>
<td>-.02</td>
<td>.11</td>
</tr>
<tr>
<td>Agreeableness → Conflict-Avoidance Goal</td>
<td>-.51**</td>
<td>.15</td>
</tr>
<tr>
<td><strong>b path</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal Goal → Overall Rating</td>
<td>.18**</td>
<td>.02</td>
</tr>
<tr>
<td>Interpersonal Goal → Rating Accuracy</td>
<td>.21**</td>
<td>.06</td>
</tr>
<tr>
<td>C-A Goal → Overall Rating</td>
<td>.01</td>
<td>.02</td>
</tr>
<tr>
<td>C-A Goal → Rating Accuracy</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td><strong>c path</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness → Overall Rating</td>
<td>-.02</td>
<td>.03</td>
</tr>
<tr>
<td>Agreeableness → Rating Accuracy</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td><strong>c’ path</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness → Overall Rating</td>
<td>-.02</td>
<td>.03</td>
</tr>
<tr>
<td>Agreeableness → Rating Accuracy</td>
<td>.11</td>
<td>.08</td>
</tr>
</tbody>
</table>

Bootstrapped mediation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>ab path</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpersonal Goal → Overall Rating</td>
<td>.00</td>
<td>.02</td>
<td>-.05</td>
<td>.04</td>
</tr>
<tr>
<td>Interpersonal Goal → Rating Accuracy</td>
<td>.00</td>
<td>.04</td>
<td>-.06</td>
<td>.05</td>
</tr>
<tr>
<td>C-A Goal → Overall Rating</td>
<td>.00</td>
<td>.02</td>
<td>-.05</td>
<td>.04</td>
</tr>
<tr>
<td>C-A Goal → Rating Accuracy</td>
<td>-.02</td>
<td>.02</td>
<td>-.08</td>
<td>.01</td>
</tr>
<tr>
<td>Total → Overall Rating</td>
<td>-.01</td>
<td>.02</td>
<td>-.06</td>
<td>.04</td>
</tr>
<tr>
<td>Total → Rating Accuracy</td>
<td>-.03</td>
<td>.04</td>
<td>-.10</td>
<td>.05</td>
</tr>
</tbody>
</table>

*Note.* N = 188. C-A Goal = conflict-avoidance goal; LL = lower limit; UL = upper limit; CI = confidence interval. ** coefficient significant at p < .01. * coefficient significant at p < .05.
Contrast between the conditional indirect effects at the three accountability conditions showed there was no significant difference in the indirect effects between the accountability conditions (see Table 3b). I conducted additional indirect effect analyses controlling for perceived accountability, but doing so did not produce meaningful changes in the results. The full moderated-mediation model with the path estimates is shown in Figure 2a for overall performance rating and in Figure 2b for rating accuracy.
<table>
<thead>
<tr>
<th>Moderator</th>
<th>Level</th>
<th>Interpersonal Goal</th>
<th>Conflict-Avoidance Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Conditional indirect effect</td>
<td>SE</td>
</tr>
<tr>
<td>Ratee</td>
<td></td>
<td>-.01</td>
<td>.02</td>
</tr>
<tr>
<td>Accountability</td>
<td>Supervisor</td>
<td>-.02</td>
<td>.03</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.01</td>
<td>.03</td>
</tr>
</tbody>
</table>

**Rating Accuracy**

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Level</th>
<th>Interpersonal Goal</th>
<th>Conflict-Avoidance Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Conditional indirect effect</td>
<td>SE</td>
</tr>
<tr>
<td>Ratee</td>
<td></td>
<td>-.01</td>
<td>.03</td>
</tr>
<tr>
<td>Accountability</td>
<td>Supervisor</td>
<td>-.02</td>
<td>.04</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.01</td>
<td>.04</td>
</tr>
</tbody>
</table>

*Note.* \( N = 188. \) LL = lower limit; UL = upper limit; CI = confidence interval. Estimates were based on 1,000 bootstrap samples. *Coefficient significant at \( p < .05. \)
Figure 2a. Mediating effect of rating goals on the relationship between rater agreeableness and overall performance rating. C-A Goal = conflict-avoidance goal.
Figure 2b. Mediating effect of rating goals on the relationship between rater agreeableness and rating accuracy. C-A Goal = conflict-avoidance goal.
Neuroticism. Table 4a shows the mediation analysis results for rater neuroticism. Results for the $a$ paths of the proposed mediation model showed that rater neuroticism significantly predicted engagement in conflict-avoidance goal, but not critical goal. In the $b$ paths of the proposed mediation model, conflict-avoidance goal significantly predicted higher overall rating and lower rating accuracy, but critical goal was not predictive of either criterion. Table 4a also shows the estimates and 95% confidence intervals for the bootstrapped total and specific indirect effects. As expected in Hypothesis 2a, results showed that rater neuroticism significantly predicted higher overall rating and lower rating accuracy through engagement in conflict-avoidance goal. The specific indirect effect through critical goal, however, was non-significant for both overall rating and rating accuracy.
Table 4a

*Simple Mediation Analysis Results for Rater Neuroticism*

<table>
<thead>
<tr>
<th>Variable</th>
<th>$b$</th>
<th>$SE$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$ path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism $\rightarrow$ Critical Goal</td>
<td>-.19</td>
<td>.10</td>
</tr>
<tr>
<td>Neuroticism $\rightarrow$ C-A Goal</td>
<td>.47**</td>
<td>.13</td>
</tr>
<tr>
<td>$b$ path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical Goal $\rightarrow$ Overall Rating</td>
<td>-.19</td>
<td>.07</td>
</tr>
<tr>
<td>Critical $\rightarrow$ Rating Accuracy</td>
<td>.00</td>
<td>.93</td>
</tr>
<tr>
<td>C-A Goal $\rightarrow$ Overall Rating</td>
<td>.06</td>
<td>.02</td>
</tr>
<tr>
<td>C-A Goal $\rightarrow$ Rating Accuracy</td>
<td>.11</td>
<td>.04</td>
</tr>
<tr>
<td>$c$ path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism $\rightarrow$ Overall Rating</td>
<td>.06</td>
<td>.03</td>
</tr>
<tr>
<td>Neuroticism $\rightarrow$ Rating Accuracy</td>
<td>.04</td>
<td>.07</td>
</tr>
<tr>
<td>$c'$ path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism $\rightarrow$ Overall Rating</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>Neuroticism $\rightarrow$ Rating Accuracy</td>
<td>-.02</td>
<td>.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bootstrapped mediation results</th>
<th>$ab$ path</th>
<th>$SE$</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Goal $\rightarrow$ Overall Rating</td>
<td>.01</td>
<td>.01</td>
<td>-.01</td>
<td>.03</td>
</tr>
<tr>
<td>Critical Goal $\rightarrow$ Rating Accuracy</td>
<td>.00</td>
<td>.01</td>
<td>-.02</td>
<td>.03</td>
</tr>
<tr>
<td>C-A Goal $\rightarrow$ Overall Rating</td>
<td>.03*</td>
<td>.01</td>
<td>.01</td>
<td>.06</td>
</tr>
<tr>
<td>C-A Goal $\rightarrow$ Rating Accuracy</td>
<td>.05*</td>
<td>.03</td>
<td>.01</td>
<td>.12</td>
</tr>
<tr>
<td>Total $\rightarrow$ Overall Rating</td>
<td>.04*</td>
<td>.01</td>
<td>.01</td>
<td>.06</td>
</tr>
<tr>
<td>Total $\rightarrow$ Rating Accuracy</td>
<td>.05*</td>
<td>.03</td>
<td>.01</td>
<td>.12</td>
</tr>
</tbody>
</table>

*Note.* $N = 188$. C-A Goal = conflict-avoidance goal; LL = lower limit; UL = upper limit; CI = confidence interval. ** coefficient significant at $p < .01$. * coefficient significant at $p < .05$. 
Contrast between the conditional indirect effects at the three accountability conditions showed there was no significant difference in the indirect effects between the accountability conditions (see Table 4b). I conducted additional analyses controlling for perceived accountability, but the results did not change. The full moderated-mediation model with the path estimates is shown in Figure 3a for overall performance rating and in Figure 3b for rating accuracy.
Table 4b  
**Moderated Mediation Analysis Results for Rater Neuroticism**

**Overall Performance**

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Level</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratee</td>
<td></td>
<td>.01</td>
<td>.01</td>
<td>-.01</td>
<td>.01</td>
<td>.03*</td>
<td>.01</td>
<td>.01</td>
<td>.08</td>
</tr>
<tr>
<td>Accountability</td>
<td>Supervisor</td>
<td>.01</td>
<td>.01</td>
<td>-.01</td>
<td>.01</td>
<td>.03*</td>
<td>.02</td>
<td>.01</td>
<td>.09</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.01</td>
<td>.01</td>
<td>-.01</td>
<td>.01</td>
<td>.03*</td>
<td>.02</td>
<td>.01</td>
<td>.08</td>
</tr>
</tbody>
</table>

**Rating Accuracy**

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Level</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratee</td>
<td></td>
<td>.00</td>
<td>.01</td>
<td>-.02</td>
<td>.01</td>
<td>.05*</td>
<td>.02</td>
<td>-.11</td>
<td>-.01</td>
</tr>
<tr>
<td>Accountability</td>
<td>Supervisor</td>
<td>.01</td>
<td>.02</td>
<td>-.02</td>
<td>.02</td>
<td>.05*</td>
<td>.03</td>
<td>-.13</td>
<td>-.01</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>.00</td>
<td>.01</td>
<td>-.01</td>
<td>.01</td>
<td>.06*</td>
<td>.03</td>
<td>-.15</td>
<td>-.01</td>
</tr>
</tbody>
</table>

*Note. N = 188. LL = lower limit; UL = upper limit; CI = confidence interval. Estimates were based on 1,000 bootstrap samples. *coefficient significant at \( p < .05 \).
Figure 3a. Mediating effect of rating goals on the relationship between rater neuroticism and overall performance rating.
Figure 3b. Mediating effect of rating goals on the relationship between rater neuroticism and rating accuracy.
Conscientiousness. Table 5a shows the mediation analysis results for rater conscientiousness. Results for the $a$ paths of the proposed mediation model showed that rater conscientiousness significantly predicted engagement in fairness goal, but it did not predict engagement in critical goal. In the $b$ paths of the proposed mediation model, engagement in critical goal significantly predicted lower overall rating and more accurate ratings. However, contrary to my expectations, engagement in fairness goal was actually predictive of higher overall rating and lower rating accuracy. Table 5a also shows the estimates and 95% confidence intervals for the bootstrapped total and specific indirect effects. The results for the indirect effect through engagement in critical goal generally showed support for Hypothesis 3. That is, rater conscientiousness was significantly related to providing more accurate ratings through engagement in critical goal. The indirect effect of rater conscientiousness through critical goal on overall rating, however, did not quite reach statistical significance. Interestingly, even though the indirect effect through equity fairness goal was statistically significant for both overall rating and rating accuracy, the results were in the opposite direction of the hypothesized effect. Although rater conscientiousness was associated with more engagement in equity fairness goal as expected, to my surprise, equity fairness goal was associated with higher overall rating and lower rating accuracy.
Table 5a

*Simple Mediation Analysis Results for Rater Conscientiousness*

<table>
<thead>
<tr>
<th>Variable</th>
<th>b</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>a path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness → Critical Goal</td>
<td>.21</td>
<td>.11</td>
</tr>
<tr>
<td>Conscientiousness → Fairness Goal</td>
<td>.49**</td>
<td>.09</td>
</tr>
<tr>
<td>b path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical Goal → Overall Rating</td>
<td>-.07*</td>
<td>.03</td>
</tr>
<tr>
<td>Critical Goal → Rating Accuracy</td>
<td>-.14**</td>
<td>.05</td>
</tr>
<tr>
<td>Fairness Goal → Overall Rating</td>
<td>.07*</td>
<td>.03</td>
</tr>
<tr>
<td>Fairness Goal → Rating Accuracy</td>
<td>.32**</td>
<td>.06</td>
</tr>
<tr>
<td>c path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness → Overall Rating</td>
<td>-.03</td>
<td>.04</td>
</tr>
<tr>
<td>Conscientiousness → Rating Accuracy</td>
<td>.02</td>
<td>.08</td>
</tr>
<tr>
<td>c' path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness → Overall Rating</td>
<td>-.06</td>
<td>.04</td>
</tr>
<tr>
<td>Conscientiousness → Rating Accuracy</td>
<td>-.11</td>
<td>.08</td>
</tr>
</tbody>
</table>

**Bootstrapped mediation results**

<table>
<thead>
<tr>
<th>Variable</th>
<th>ab path</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Goal → Overall Rating</td>
<td>-.01</td>
<td>.01</td>
<td>-.04</td>
<td>&gt;.00</td>
</tr>
<tr>
<td>Critical Goal → Rating Accuracy</td>
<td>-.03*</td>
<td>.02</td>
<td>-.09</td>
<td>&lt;.00</td>
</tr>
<tr>
<td>Fairness Goal → Overall Rating</td>
<td>.04*</td>
<td>.02</td>
<td>&lt;.01</td>
<td>.08</td>
</tr>
<tr>
<td>Fairness Goal → Rating Accuracy</td>
<td>.16*</td>
<td>.04</td>
<td>.08</td>
<td>.25</td>
</tr>
<tr>
<td>Total → Overall Rating</td>
<td>.02</td>
<td>.02</td>
<td>-.01</td>
<td>.06</td>
</tr>
<tr>
<td>Total → Rating Accuracy</td>
<td>.13*</td>
<td>.04</td>
<td>.06</td>
<td>.22</td>
</tr>
</tbody>
</table>

*Note. N = 188. C-A Goal = conflict-avoidance goal; LL = lower limit; UL = upper limit; CI = confidence interval. ** coefficient significant at p < .01. * coefficient significant at p < .05.*
Contrast between the conditional indirect effects at the three accountability conditions showed there was no significant difference in the indirect effects between the accountability conditions (see Table 5b). I conducted additional analyses controlling for perceived accountability, but it did not produce meaningful changes in the results. The full moderated-mediation model with the path estimates is shown in Figure 4a for overall performance rating and in Figure 4b for rating accuracy.
<table>
<thead>
<tr>
<th>Moderator</th>
<th>Level</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratee</td>
<td></td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.02*</td>
<td>0.02</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>Accountability</td>
<td>Supervisor</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.04*</td>
<td>0.02</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.04*</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Moderator</th>
<th>Level</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
<th>Conditional indirect effect</th>
<th>SE</th>
<th>LL 95% CI</th>
<th>UL 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratee</td>
<td></td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.11*</td>
<td>0.04</td>
<td>0.03</td>
<td>0.19</td>
</tr>
<tr>
<td>Accountability</td>
<td>Supervisor</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.5</td>
<td>0.2</td>
<td>0.15*</td>
<td>0.05</td>
<td>0.05</td>
<td>0.21</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td>0.00</td>
<td>0.01</td>
<td>-0.2</td>
<td>0.2</td>
<td>0.08*</td>
<td>0.04</td>
<td>0.02</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Note. N = 188. LL = lower limit; UL = upper limit; CI = confidence interval. Estimates were based on 1,000 bootstrap samples. * coefficient significant at p < .05.
Figure 4a. Mediating effect of rating goals on the relationship between rater conscientiousness and overall performance.
Figure 4b. Mediating effect of rating goals on the relationship between rater conscientiousness and rating accuracy.
Accountability. Mediation analysis results for accountability context did not show support for Hypothesis 4a or Hypothesis 4b. Contrary to my expectations, raters in the supervisor accountability condition did not engage in more equity fairness goal ($b = .60, t = .92, p = .36$) or critical goal ($b = .34, t = .45, p = .66$) than raters in the baseline condition. Similarly, raters in the ratee accountability condition did not engage in more interpersonal goal ($b = .57, t = .83, p = .41$) or conflict-avoidance goal ($b = 1.13, t = 1.16, p = .25$) than raters in the baseline condition. There was also non-significant difference in overall rating and rating accuracy between raters in the accountability conditions and raters in the baseline condition, even after controlling for perceived accountability.

Mean Overall Performance Rating and Rating Accuracy Comparisons

Rater personality. Table 6 shows the mean overall performance rating and rating accuracy comparison test results for each rater personality traits. Contrary to previous study results that have shown that agreeable raters tend to provide more lenient and inaccurate ratings, I found no significant difference in mean overall rating or rating accuracy between high vs. low agreeable raters. Similarly, contrary to previous study results that have shown that conscientious raters tend to provide lower and more accurate ratings, there was non-significant difference in mean overall rating or rating accuracy between high vs. low conscientious raters. I conducted the analyses again controlling for accountability condition and perceived accountability to examine whether the results were confounded by the difference in the experimental condition. This however, did not significantly influence the results. In line with the mediation analysis results for rater neuroticism, high neurotic raters provided significantly more lenient ratings than low neurotic raters, $t(186) = 2.01, p = .04$, $d$
=.31, although I did not find a significance difference in rating accuracy between high vs. low neurotic raters.

**Accountability.** Table 6 also shows the between-accountability conditions mean comparison test results. Results showed that participants in the supervisor accountability condition provided more accurate ratings than participants in the baseline condition \((t = 2.44, p = .02, d = .45)\), providing partial support for Hypothesis 4b. Although participants in the supervisor accountability were also more accurate than participants in the ratee accountability condition, this difference was not statistically significant. I also hypothesized that participants in the ratee accountability condition would provide more inaccurate ratings than participants in the baseline condition (Hypothesis 4a), but the data did not show support for this hypothesis.

Contrary to my expectations, I found non-significant difference in overall rating between raters in the supervisor accountability condition and baseline condition, and between raters in the ratee accountability condition and baseline condition. The mean difference between raters in the supervisor accountability condition and raters in the ratee accountability condition was also non-significant. Controlling for participants’ level of perceived accountability did not produce meaningful difference in any of the mean comparison test results.
Table 6

*Mean Difference between Personality Trait and Accountability Conditions*

<table>
<thead>
<tr>
<th>Variable</th>
<th>$M_{Rating}$ (SD)</th>
<th>$t$</th>
<th>$p$</th>
<th>$d$</th>
<th>$M_{Accuracy}$ (SD)</th>
<th>$t$</th>
<th>$p$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agreeableness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>4.61 (1.46)</td>
<td>.77</td>
<td>.44</td>
<td>.12</td>
<td>2.82 (2.65)</td>
<td>.73</td>
<td>.47</td>
<td>.11</td>
</tr>
<tr>
<td>Low</td>
<td>4.77 (1.25)</td>
<td></td>
<td></td>
<td></td>
<td>2.52 (2.95)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Neuroticism</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>4.84 (1.07)</td>
<td>2.01*</td>
<td>.04</td>
<td>.31</td>
<td>2.73 (2.77)</td>
<td>.24</td>
<td>.81</td>
<td>.03</td>
</tr>
<tr>
<td>Low</td>
<td>4.49 (1.17)</td>
<td></td>
<td></td>
<td></td>
<td>2.64 (2.81)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conscientiousness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>4.56 (1.42)</td>
<td>1.55</td>
<td>.12</td>
<td>.24</td>
<td>2.62 (2.67)</td>
<td>.42</td>
<td>.68</td>
<td>.06</td>
</tr>
<tr>
<td>Low</td>
<td>4.88 (1.27)</td>
<td></td>
<td></td>
<td></td>
<td>2.80 (2.98)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Accountability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratee</td>
<td>4.64 (1.37)</td>
<td></td>
<td></td>
<td></td>
<td>2.60 (2.48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supervisor</td>
<td>4.56 (1.22)</td>
<td>.64</td>
<td>.53</td>
<td>.01</td>
<td>2.07 (2.25)</td>
<td>3.32*</td>
<td>.04</td>
<td>.03</td>
</tr>
<tr>
<td>Baseline</td>
<td>4.83 (1.51)</td>
<td></td>
<td></td>
<td></td>
<td>3.34 (3.35)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. $N = 188$. $M_{Rating}$ = mean overall performance rating; $M_{Accuracy}$ = mean rating accuracy; $SD$ = standard deviation. * $F$ value significant at $p < .05$. 


Agreeableness × Accountability interaction. I conducted a 2 (rater agreeableness; high vs. low) × 2 (accountability context; ratee vs. baseline) between-subjects ANOVA to test the hypothesized interactive effect of rater agreeableness and accountability context on overall performance rating and rating accuracy. The results, however, did not show support for the hypothesized interaction for both overall performance rating and rating accuracy (Hypothesis 6a; see Table 7). Controlling for perceived accountability did not produce any meaningful difference in the results.
Table 7

*Rater Agreeableness × Accountability Context ANOVA results*

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>F</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DV: Overall Performance Rating</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.08</td>
<td>.00</td>
</tr>
<tr>
<td>Accountability</td>
<td>.54</td>
<td>.00</td>
</tr>
<tr>
<td>Two-way interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness × Accountability</td>
<td>.05</td>
<td>.00</td>
</tr>
<tr>
<td><strong>DV: Rating Accuracy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.18</td>
<td>.00</td>
</tr>
<tr>
<td>Accountability</td>
<td>2.15</td>
<td>.02</td>
</tr>
<tr>
<td>Two-way interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agreeableness × Accountability</td>
<td>.05</td>
<td>.00</td>
</tr>
</tbody>
</table>

*Note. N = 129. Accountability main effect compared ratings between raters in the ratee accountability condition vs. raters in the baseline condition.*
Neuroticism × Accountability interaction. I conducted a 2 (rater neuroticism; high vs. low) × 3 (accountability context; ratee vs. supervisor vs. baseline) between-subjects ANOVA to test the hypothesized interactive effect of rater neuroticism and accountability context on overall performance rating and rating accuracy. As indicated in the mean comparison test results mentioned above, high neurotic raters provided significantly higher overall ratings than low neurotic raters. The results, however, did not show support for the hypothesized interaction for both overall performance rating and rating accuracy (Hypothesis 6b; see Table 8).
Table 8
*Rater Neuroticism × Accountability Context ANOVA results*

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>$F$</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV: Overall Performance Rating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>3.92*</td>
<td>.02</td>
</tr>
<tr>
<td>Accountability</td>
<td>.53</td>
<td>.01</td>
</tr>
<tr>
<td>Two-way interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism ×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accountability</td>
<td>2.76</td>
<td>.02</td>
</tr>
<tr>
<td>DV: Rating Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>.01</td>
<td>.00</td>
</tr>
<tr>
<td>Accountability</td>
<td>3.52*</td>
<td>.04</td>
</tr>
<tr>
<td>Two-way interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism x Accountability</td>
<td>.42</td>
<td>.01</td>
</tr>
</tbody>
</table>

*Note.* $N = 187$. *F* value significant at $p < .05$. 

For exploratory purposes, I conducted a two-way ANOVA with rater neuroticism (high vs. low) and the two accountability conditions (ratee vs. supervisor) as the two factors. For overall performance rating, there was a significant neuroticism main effect, $F(1, 122) = 9.80, p = .02$, partial $\eta^2 = .05$ (high neurotic raters provided higher ratings), but a non-significant accountability main effect, and neuroticism by accountability interaction. For rating accuracy, there was a significant rater neuroticism main effect, $F(1, 122) = 4.05, p = .04$, partial $\eta^2 = .02$ (high neurotic raters significantly less accurate), a significant accountability main effect, $F(1, 122) = 4.23, p = .03$, partial $\eta^2 = .03$ (raters in the ratee accountability condition significantly less accurate), and a significant neuroticism by accountability interaction, $F(1, 122) = 4.11, p = .04$, partial $\eta^2 = .02$. To interpret this interaction, I conducted a series of post hoc pairwise comparisons between high and low neurotic raters within each accountability condition, and also between the two accountability conditions within each rater neuroticism group ($p$ values were adjusted using Bonferroni correction to control for family-wise Type I error rate). The results showed that there was a significant difference in rating accuracy between high vs. low neurotic raters in the ratee accountability condition, $F(1, 122) = 7.55, p = .03$, partial $\eta^2 = .05$ (high neurotic raters were significantly less accurate), but not in the supervisor accountability condition. In fact, low neurotic raters in the ratee accountability condition and the supervisor accountability condition showed similar level of rating accuracy, but high neurotic raters in the ratee accountability condition showed significantly lower rating accuracy than high neurotic raters in the supervisor accountability condition, $F(1, 122) = 9.02, p < .01$, partial $\eta^2 = .01$.

**Conscientiousness ✕ Accountability interaction.** I conducted a 2 (rater conscientiousness; high vs. low) ✕ 2 (accountability context; supervisor vs. baseline)
between-subjects ANOVA to test the hypothesized interactive effect of rater conscientiousness and accountability context on overall performance rating and rating accuracy. The results, however, did not show support for the hypothesized interaction for both overall performance rating and rating accuracy (Hypothesis 6c; see Table 9). Although the marginally significant conscientiousness main effect for overall performance rating ($p = .056$) emerged as significant when controlled for perceived accountability, $F (1, 123) = 4.89, p = .03$, partial $\eta^2 = .04$ (high conscientious raters provided significantly lower ratings), it did not influence the results for the hypothesized rater conscientiousness by accountability interactions.
<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>DV: Overall Performance Rating</th>
<th>DV: Rating Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$</td>
<td>Partial $\eta^2$</td>
</tr>
<tr>
<td>Main effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>3.71</td>
<td>.03</td>
</tr>
<tr>
<td>Accountability</td>
<td>1.10</td>
<td>.01</td>
</tr>
<tr>
<td>Two-way interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accountability</td>
<td>.33</td>
<td>.00</td>
</tr>
</tbody>
</table>

**Note.** $N = 124$. Accountability main effect compared ratings between raters in the supervisor accountability condition vs. raters in the baseline condition. ** $F$ value significant at $p < .01$. 
CHAPTER 6
Discussion

Theoretical models of rater motivation (e.g., Harris, 1994; Murphy & Cleveland, 1991, 1995; Spence & Keeping, 2013) proposed that raters assess the benefits and consequences of different rating behaviors in a performance appraisal context, and pursue a rating behavior that is judged to be the most favorable. These models proposed that raters’ perception of favorability of different rating behaviors are affected by either situational or personal factors. Following this theoretical framework, I developed hypotheses on how different accountability situations and rater personality traits might jointly affect rater motivation in performance appraisal context and performance ratings, and then empirically tested these hypotheses. I found several noteworthy results that warrant further discussion.

Rater Personality

Neuroticism. One of the main contributions of the current study to the performance appraisal literature was demonstrating the effect of rater neuroticism on rater motivation in performance appraisal and performance ratings. As hypothesized, the results showed that rater neuroticism was positively associated with motivation to avoid providing ratings that conflict with an audience’s presumed preference, and this conflict-avoidance motivation was predictive of higher overall rating and lower rating accuracy in general. Neurotic raters’ preference to avoid conflict with the audience of the appraisal process provides an explanation for the significant interaction between rater neuroticism (high vs. low) and accountability condition (ratee vs. supervisor). That is, with the assumption that raters held accountable to ratees generally account for their ratings by providing positively distorted ratings, and raters held accountable
to authorities generally account for their behavior by providing accurate ratings (Mero et al., 2007), high neurotic raters provided ratings that were in line with the preference of the rating audience (less accurate ratings in the ratee accountability condition; more accurate ratings in the supervisor accountability condition), whereas low neurotic raters’ ratings were not significantly influenced by the rating audience.

It is also notable that high neurotic raters were the only group for which I replicated (at least partially) the general effect of accountability pressure found in the previous literature (inflated, less accurate ratings when held accountable to ratees vs. lower, more accurate ratings when held accountable to supervisors; Mero et al., 2007). Although these results should be interpreted carefully given some of the important limitations in the current study (which I will discuss later), the effect of rater neuroticism on performance ratings through engagement in conflict-avoidance goal, and neuroticism by accountability interaction results for rating accuracy suggest that highly neurotic raters may be especially sensitive to the contextual demands of a performance appraisal situation, making them prone to rating distortion in the face of social pressure to provide ratings in a predisposed manner as a consequence of their motivation to avoid facing the stressful consequences that may be associated with going against the contextual demands (e.g., presumed preferences that the audience have about the ratings).

**Agreeableness.** The current study results generally did not show support for the hypothesized effect of rater agreeableness on rating motivation and performance ratings. In fact, some of the findings were in direct opposite direction of the hypothesized effect. The most surprising finding was the negative relationship between rater agreeableness and engagement in conflict-avoidance goal. Both theory and previous empirical evidence suggested that agreeable individuals tend to be
generous, and motivated to avoid conflict with others, which is why in the context of performance appraisal, I expected that agreeable raters would be motivated to avoid conflict with whomever they evaluate or whomever is overlooking the evaluation process (and even more so when they are held accountable to an audience). The results, however, showed a strong negative effect of rater agreeableness on engagement in conflict-avoidance goal in the hypothesized indirect effect model. Also, contrary to the previous studies that have shown that agreeable raters generally provide more lenient and less accurate ratings (both in conditions of relatively high and low rater accountability; Bernardin et al., 2000; Bernardin et al., 2009), the current study results did not show significant differences in overall rating level and rating accuracy between raters high vs. low in agreeableness. Furthermore, I did not find support for the agreeableness × accountability context interaction found in previous studies (e.g., Roch et al., 2005).

Exploring the data more carefully, I noticed that there was a strong positive correlation between participants’ scores on agreeableness and conscientiousness beyond the low correlation between these two variables that has been found in the literature. For example, Chang, Connelly, and Geeza (2012) conducted a multitrait-multimethod (MTMM) meta-analysis to estimate trait correlations among the Big Five factors found a modest correlation between agreeableness and conscientiousness in the monomethod correlations ($\rho = .24$; single informant ratings of a target person’s level of agreeableness and conscientiousness). However, the correlation value dissipated in the heteromethod correlations ($\rho = .01$ in the self-informant ratings correlation; $\rho = .06$ in the informant-informant ratings correlation), suggesting that the modest correlation found in the monomethod correlations is inflated due to common method variance. Other primary studies have also found low correlation between
agreeableness and conscientiousness (e.g., Witt, Burke, Barrick, & Mount, 2002). The unusually high correlation between agreeableness and conscientiousness found in the current study could indicate that the effect of rater conscientiousness may have confounded the effect of rater agreeableness on rating motivation, and performance ratings (and vice-versa). I re-examined the proposed mediation effect model and the mean difference tests for rater agreeableness with rater conscientiousness as a covariate (to control for the effect of rater conscientiousness on the hypothesized effects of rater agreeableness; e.g., relationship between rater agreeableness and engagement in conflict-avoidance goal). This however, did not produce meaningful difference in the study results. Controlling for other variables possibly confounding the effect of rater agreeableness (accountability condition, perceived accountability) also did not significantly influence the results. Based on Bernardin and colleagues’ (Bernardin et al., 2000; Bernardin et al., 2009) finding that raters with high agreeableness and low conscientiousness provided the most lenient ratings, I also compared the difference in overall rating and rating accuracy between raters who theoretically have a personality profile with the most lenient tendencies (e.g., high agreeableness × low conscientiousness; high neuroticism × low conscientiousness) and raters who theoretically have a personality profile with the least lenient tendencies (low agreeableness × high conscientiousness; low neuroticism × high conscientiousness), but I did not find any meaningful results.

A notable potential explanation for these counterintuitive study results is that the effect of rater agreeableness may have been driven by a substantively higher-order factor, particularly social desirability. Participants were told at the beginning of the experiment that the goal of this study was to examine how raters conduct performance appraisal. As experienced supervisors, participants may have been motivated to
respond to the questionnaires in a manner that represents how supervisors ought to evaluate employees in a performance appraisal context in order to present themselves as responsible raters.

Data showed some evidence for this possibility. In addition to the unusually high correlation between participants’ scores on agreeableness and conscientiousness, these two personality variables showed a very similar pattern of correlations with other main study variables. The high correlation between rater agreeableness and conscientiousness, and similar pattern of relationships with the other study variables indicate that there may have been a substantively higher-order psychological factor that motivated raters to endorse agreeableness and conscientiousness personality items (both positive personality traits). Descriptive statistics for the main study variables suggest that participants may have responded to the study questionnaires in a socially desirable manner. Specifically, participants reported being very high on desirable personality traits (agreeableness and conscientiousness), whereas they reported being relatively low on neuroticism, which is a less socially desirable personality trait. Similarly, participants generally endorsed engaging in rating goals that are seemingly associated with providing accurate, honest performance ratings (critical goal and equity fairness goal), whereas they reported relatively lower engagement in conflict-avoidance goal, which is generally viewed as an ineffective appraisal behavior (these patterns were observed within each accountability condition). Taken together, the counterintuitive relationship between rater agreeableness and conflict-avoidance goal might have been found because participants’ motivation to provide socially desirable responses was affecting this relationship. That is, participants’ motivation to endorse desirable personality trait items (agreeableness and conscientiousness) may have also motivated them to
respond negatively to undesirable performance appraisal behavior items (conflict-avoidance goal).

**Conscientiousness.** The current study results generally supported the hypothesis that rater conscientiousness would predict lower overall performance rating through engagement in critical goal, but contrary to my expectation, the indirect effect of rater conscientiousness through engagement in equity fairness goal was significantly associated with greater leniency and inaccurate ratings. The current study results also did not replicate previous study results that have shown that conscientious raters tend to provide lower and more accurate ratings (e.g., Bernardin et al., 2009), and that they tend to be less influenced by the accountability context relative to raters low in conscientiousness (e.g., Roch et al., 2005). Similar to controlling for the effect of rater conscientiousness in testing the effect of rater agreeableness on performance ratings, I re-examined the study results for rater conscientiousness controlling for potential confounding effect that rater agreeableness may have had on the study results. However, this did not significantly influence the analysis outcomes.

The counterintuitive effect of conscientiousness on performance ratings through engagement in equity fairness goal might be because participants knew that the performance appraisal task was for research purposes. Previous studies have shown that raters are generally less motivated, less attentive, and less sensitive to performance information in performance appraisals conducted for research purposes than in performance appraisals conducted for administrative purposes (Murphy & Cleveland, 1995). With low perceived importance of performance appraisal, participants may have perceived the ratee’s performance in a more positive manner
and forego exercising the effort that is required to carefully attend to, and critically evaluate specific performance behaviors.

**Accountability.** As briefly discussed above, the null effect of accountability context generally found in the current study may be attributed to the relatively weak accountability manipulation. For practical reasons (increasing participation) I could not ask participants to identify themselves. This was an important limitation relative to the previous studies on performance appraisal, many of which manipulated accountability by identifiability (e.g., Roch & McNall, 2007). With respect to Schlenker’s accountability pyramid (which proposed that an individual’s perceived accountability for his/her behavior increases as the link between an individual’s identity and an individual’s decision is stronger; Schlenker, 1986), anonymity of the participants is likely to have greatly undermined the effect of the accountability manipulation in the current study.

Additionally, accountability theory posits that when one is held accountable, one is likely to behave in a manner that is pleasing to the demands of the audience, which allows the accountable agent to maintain the approval of the audience and assume the least amount of liability. In my experiment, however, accountability manipulation is unlikely to have imposed contextual demands upon the participants that are as strong as the demands that are placed on the raters in an actual performance appraisal setting. There may be various reasons for this, but most importantly, (1) the participants knew that they would not have to actually face the ratee or the experimenters, and (2) ratings were not tied to any practical consequences on the part of the ratee, the participant, or the experimenters. These factors are likely to have made the participants indifferent about the consequences of the ratings that they provide (or participants perceived that there were no consequences), which in
turn may have prompted the participants to be negligent to the demands that come with being held accountable. Also, the experience of my sample may have caused such negligence of the accountability manipulation to be stronger because they understand what the actual contextual demands in practice are, which are probably far from the demands that are placed on the participants through the accountability manipulation.

**Theoretical Implications**

The defining feature of the goal-directed performance appraisal perspective is its emphasis on the rater’s choice of intentions in the performance appraisal process. In this framework is, in addition to ratee’s actual job performance, a host of rater factors and contextual factors are expected to prompt raters to assume different rating goals in evaluating a ratee’s job performance, and its specific rating goals that raters pursue as a function of these factors that largely predict the ratings that raters provide (Murphy, 2008; Murphy & Cleveland, 1991, 1995). The current study results provide partial support for the goal-directed performance appraisal perspective, for which the literature has remained largely theoretical (e.g., Harris, 1994; Murphy & Cleveland, 1995; Spence & Keeping, 2011, 2013). Some studies have shown empirical support for the goal-directed performance appraisal perspective (e.g., Murphy et al., 2004; Wong & Kwong, 2007), but the current work extends the previous literature by testing the broader theoretical mediation model conception of the effect that both contextual (accountability), and personal (rater personality) have on performance ratings through their influence on how raters choose to approach performance appraisal.

Additionally, research on the effect of rater personality on performance ratings has primarily focused on agreeableness and conscientiousness. Results from the
current study showed that rater neuroticism is another personality that significantly influences how raters approach performance appraisal and how they rate their employees. It is noteworthy that I found some reliable evidence for the effect of rater neuroticism on performance appraisal despite some crucial limitations in the current study, which I will discuss later.

**Practical Implications**

Ideally, organizations would like to avoid having raters who are prone to rating distortion from giving performance ratings, but this might be a difficult goal to achieve considering the practical hurdles that are likely to be involved (e.g., determining the criteria for rater selection decision; selecting supervisors for effective appraisal vs. selecting supervisors for effective managerial performance). More realistically, organizations should consider putting in place interventions for raters who are identified as more likely to give problematic performance ratings. Results from the current study suggest that those who feel more anxious (individuals who score high in neuroticism) are especially prone to pressure from performance appraisal context, and as a consequence, elect to provide performance ratings that help them assume the least amount of liability, which in turn produces distorted in performance ratings. To alleviate this effect, organizations might modify their performance appraisal system to make it difficult to pinpoint the person who provided the ratings (e.g., performance appraisal by committee; double-blind performance appraisal). This might alleviate raters’ anxiety about the possibility of facing the negative consequences of providing accurate, honest ratings.

At a more general level, organizations should cultivate an environment that is conducive to exchanging critical and honest performance feedback information with less concern towards interpersonal political considerations that could negatively
influence the effectiveness of the performance appraisal process. Researchers have pointed out that one of the problems of formal performance appraisal system is that it has to summarize an employee’s vast range of job performance episodes over an extended time period into a single number (or a single number per dimension of job performance). This could cause employees to perceive performance ratings as being disconnected from their day-to-day activities, and that the ratings do not accurately represent their job performance, which in turn could result in negative reactions to performance ratings, especially when the ratings are not as high as the ratees expected.

To this effect, instead of relying only on an intermittent formal performance appraisal, organizations might consider adopting a more informal, continuous performance management system where employees constantly observe, evaluate, and provide real-time feedback on each other’s on-the-job performance behaviors (Pulakos, Hanson, Arad, & Moye, in press). Not only is continuous feedback more effective in creating real-time improvement in job performance behaviors than intermittent formal feedback (Cederblom, 1982), the track record of employees’ ability to respond to continuous feedback information would make it easier for raters to justify any managerial decisions that are based on that record (e.g., promotion), and for employees to accept those decisions.

**Limitations and Future Research**

There are several limitations in the current study, many of which offer the possibility of areas for future research. First and foremost, the appraisal task in this study was an online experimental exercise, which was far from being representative of the complex nature of an organizational social environment that complicate how raters approach performance appraisal. My initial thinking was that gathering data from actual supervisors would make the study results more generalizable to practice.
However, this meant that for practical reasons, I would have to conduct the study online, which made it extremely difficult to design an experiment that contained the realistic social complexities in an organizational environment, most notably a realistic accountability manipulation. In fact, social factors that complicate performance appraisal in organizations were captured quite well in previous studies that used a college student sample. For example, participants in the experimental group in these studies were asked to account for their ratings to their course instructor (e.g., Roch & McNall, 2007), or peers or group members in their class (e.g., Roch et al., 2005). Participants in these studies are likely to have perceived realistic interpersonal pressure in the performance appraisal task because providing performance ratings that the ratee may not be satisfied with was associated with actual practical consequences on the part of the participants. For example, in Roch and McNall’s (2007) study, students held accountable to the instructor for their instructor evaluation would have taken into consideration the fact that the instructor, who is responsible for providing the students’ grades, would be able to identify who provided the ratings. However, it is notable that I was able to find some support for my study hypotheses, which emphasize the important effect that accountability context and rater personality have on performance ratings.

Second, the rating goal questionnaire used in the current study did not capture the full range of rating goals that raters pursue in performance appraisal process in practice. Although representing the full spectrum of rating goals was not the intention of the current study, it is highly likely that I have overlooked some important rating intentions that meaningfully influence how raters approach performance appraisal. In particular, omission of the developmental goals is an important aspect of this study limitation. Based on the data that is available, it is impossible to infer the exact nature
of the participants’ rating intention behind their reporting of engagement in critical goal. Based on previous research findings that have shown that the purpose of the appraisal significantly influences the accuracy and the level of ratings that the raters assign (Harris, 1994; Murphy & Cleveland, 1991), future research should examine how the purpose of the appraisal affects raters’ intentions, and how they subsequently influence the ratings that they provide.

Third, there was some discrepancy between what the accountability construct theoretically represents, and how perception of accountability was measured with the perceived accountability items. An important aspect of the accountability construct is that it should pressure the individuals held accountable to attend to the audience’s prescribed standards, and prompt them to behave in a manner that reflects the demands of the audience. The perceived accountability items however, were worded in a manner that made it difficult to infer if participants’ scores on perceived accountability reflects accountability to the demands of the audience, or quite simply their confidence in being able to answer for their ratings from their own personal standard with little regard for the demands of the audience.

Fourth, participants in the current study may not have been given enough information to be able to accurately evaluate the ratee’s performance in the experimental task. That is, the SMEs who produced the true score for the ratee in the current study evaluated his performance according to specific performance standards that are clearly laid out (see Kudisch, 2009 for methods of scoring assessment center performance). Although the ratee in the current experiment did display performance behaviors that are obviously ineffective (e.g., agreeing to terminate an employee without sufficient basis), participants may not have had enough information about the context of the performance episode or information about the criteria for
discriminating between effective and ineffective performance to make a meaningful judgment about the ratee’s performance.

Lastly, construct validity of the rating goal questionnaire may be of concern. Although CFA showed adequate model fit, and rater personality traits and rating goals showed correlation patterns that made theoretical sense for the most part, more evidence is needed to ensure that the rating goal questionnaire items accurately represents the theoretical definitions of the rating goals examined in the current study.

Related to the previous point, researchers should consider a more effective method of measuring rater intentions in the performance appraisal process. Spence and Keeping (2011) expressed that studying rater intentions is difficult because raters are often unwilling to admit to engaging in biased rating motivation. I mentioned earlier that descriptive statistics and correlation patterns among the study variables suggest that participants may have tried to provide socially desirable answers to the study items. These indicate that directly asking raters identify how inaccurately they approached performance appraisal may not work, even under a research setting where there is minimal cost for providing honest responses to the questionnaires.

Conclusion

Although the criterion problem has been an important issue in the management psychology for a long period of time, research in this area still has many questions to answer. With the increasing prevalence of new performance appraisal practices in organizations, research needs to keep pace with practice to offer guidelines that help organizations make better use of their performance appraisal system.

The current study results showed that individual differences in rater personality are important predictors of performance ratings, and that rater personality
indirectly predicts performance ratings through rating motivations that raters choose to pursue in the performance appraisal process. However, more research is clearly needed to delineate the effect that multitude of situational and personal factors that might influence how raters approach performance appraisal, and the ratings that they provide.
References


Appendix

Rating Goal Scale

Instructions

On a scale ranging from 1 (strongly disagree) to 5 (strongly agree), please rate the extent to which you engaged in each of these goals when you were evaluating the manager’s performance.

Interpersonal

1. I was motivated to be kind to the manager.
2. I was motivated to rate the manager’s performance in a manner that would help maintain his self-esteem.
3. I was motivated to be polite to the manager.
4. I was motivated to provide ratings that the manager would deem as pleasant to receive.
5. I was motivated to evaluate the manager in a manner that would help improve his confidence.

Conflict-Avoidance

1. I was motivated to provide ratings that would help me avoid facing trouble regarding my ratings.
2. I was motivated to provide ratings that would help me avoid questions from the person reviewing my ratings.
3. I was motivated to provide ratings that would help me avoid conflict with the person reviewing my ratings.
4. I was motivated to distort my ratings to avoid conflict with the person reviewing my ratings.
5. I was motivated to provide ratings that would help me avoid facing confrontations about the ratings that I provided.

**Critical**

1. I was motivated to rate the manager in a manner that accurately indicates what was done poorly.
2. I was motivated to provide ratings that accurately reflect the manager’s deficiencies in performance.
3. I was motivated to accurately point out the manager’s deficiencies.
4. I was motivated to provide ratings that accurately reflect where the manager is falling short in terms of performance.
5. I was motivated to provide ratings that accurately reflect the manager’s behaviors of things that he did wrong.

**Equity Fairness**

1. I was motivated to rate the manager’s performance in a manner that reflects his performance.
2. I was motivated to rate the manager’s performance in such a way that he will realize it is based upon his performance.
3. I was motivated to provide performance ratings that objectively represent the manager’s performance.
4. I was motivated to provide performance ratings that the manager deserved.
5. I was motivated to uphold integrity as a fair rater.