2014

Rater Characteristics in Performance Evaluation Accuracy

Shotaro Hakoyama
Wright State University

RATER CHARACTERISTICS IN PERFORMANCE EVALUATION ACCURACY

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science

By

SHOTARO HAKOYAMA
B.S., Central Michigan University, 2010

2014
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

DATE: FEBRUARY 28, 2014

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Shotar Hakoyama ENTITLED Rater Characteristic in Performance Evaluation Accuracy BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

David LaHuis, Ph.D.
Thesis Director

Scott N. J. Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Committee on
Final Examination

David LaHuis, Ph.D.

Gary Burns, Ph.D.

Nathan Bowling, Ph.D.

Robert Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

ABSTRACT

Hakoyama, Shotaro. M.S. Department of Psychology, Wright State University, 2014. Rater Characteristics in Performance Evaluation Accuracy.

The current study examined how various rater-level variables are related to performance ratings. Student participants watched a series of video clips depicting high, medium, and low levels of employee work performance. The participants then rated the performance of the employee depicted in the video on a graphic rating scale. The rater level variables examined for the current study included rater goals, cognitive ability, conscientiousness, and agreeableness. The results indicated that focusing the on the strength of the employee (strength goal) was associated with rating elevation and that rater conscientiousness was associated with rating deflation. Strength goal, conscientiousness, agreeableness, and cognitive ability were all associated with rating accuracy. Theoretical and practical implications are discussed.

Keywords: performance evaluation, rater goals, personality, conscientiousness, agreeableness, cognitive ability, accuracy, multilevel random coefficient model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Rater Characteristics in Performance Evaluation Accuracy

Performance evaluation is an important topic in organizational research because organizations base personnel decisions on the outcome of such evaluations. Failing to provide accurate performance evaluations might result in erroneous judgments regarding employee promotions, demotions, and terminations, which can then affect the effectiveness of the organization. Historically, much of performance evaluation literature has focused on appraisal format and rater cognition (e.g., Bretz, Milkovich, & Read, 1992; Cardy & Dobbins, 1986; Hauenstein, 1992; Kraiger & Ford, 1985; Sinclair, 1988). However, more recent studies have suggested that individual rater characteristics, such as rater goals and personality, might have significant effects on performance evaluations (e.g., Kane, Bernardin, Villanova, & Peyrefitte, 1995; Murphy, Cleveland, Skattebo, & Kinney, 2004; Spence & Keeping, 2010). Whereas studies examining rater characteristics as a source of error provided a new perspective on performance appraisal research, these studies have focused primarily on how these characteristics elevate/deflate the ratings without necessarily linking the ratings to objective performance levels. In the current study, I will address this limitation of previous research by assessing how rater characteristics affect the accuracy of ratings. In the present study, I define accuracy as the ability of raters to distinguish among different performance levels.

Rater Goals

Murphy et al. (2004) examined how rater goals influence performance ratings. They hypothesized that rating inaccuracies, rather than being unintentional distortions,

were the result of conscious decision making processes and that rater goals determined degree and patterns in which the ratings were affected. For instance, a rater might be trying to maintain harmony within the work group. If a rater were to focus on this particular goal, then it would be logical for this rater to provide uniform ratings to all the subordinates and avoid giving extremely high or low scores to avoid conflict and hostility within the workgroup.

Another example of how a rater goal could affect the performance evaluation is if a rater were interested in keeping the subordinates motivated for the job. If a rater were to focus on this goal, then he or she may avoid giving low ratings and may provide generally positive feedback (Murphy et al., 2004).

Murphy et al. (2004) used teacher evaluations by the students to evaluate whether the students (the raters) followed particular goals when they evaluated the teachers. Along with the performance evaluations, they asked the raters 19 questions regarding the thought process they went through while they completed the performance evaluations of the teachers. The result indicated that rater goals explain significant amount of variance in the performance evaluations. In addition, Murphy et al. conducted a factor analysis of the questionnaire and identified four specific goals that the raters used to evaluate the teachers. The goals they identified are: identifying weakness of the ratee, identifying the strength of the ratee, being fair to the ratee, and being motivating to the ratee (Murphy et al., 2004).

The findings of Murphy et al. (2004) were meaningful in that they identified distinct rater goals, however, Murphy et al.'s study had a few shortcomings. For instance, their study had only one ratee, thus not allowing each rater to provide multiple ratings. Having only one ratee is problematic because it makes it impossible for the researchers to differentiate between mean ratings and discriminability of the ratings. Mean ratings refers to the average level of all the ratings a rater provides. Discriminability refers to the variability of the ratings a rater provides. Because there was only one ratee, Murphy et al.'s study did not allow the researcher to observe both mean ratings and discriminability. In addition, Murphy et al's study was also correlational in nature with no experimental manipulation resulting in weak support for causal relationship between rater goals and ratings. Finally, lack of control in Murphy's study made it difficult to observe different influences of rater goals (Wong & Kwong, 2007).

Wong and Kwong (2007) addressed these issues in an experimental field study using student samples. The students were asked to give the group project members peer ratings, which were then used to adjust the grades for class credit. Wong and Kwong used two different time points (mid-semester and the end of the semester) and four different goals (identification, harmony, fairness, and motivating) for the study. In the identification goal condition, the raters were asked to identify the strengths and weaknesses of the ratee. In the harmony goal condition, the raters were asked to maintain harmony within the group. In the fairness goal condition, the raters were instructed to give accurate and fair ratings for the group members. In the motivating goal condition,

the raters were instructed to use ratings to motivate the group members. Because it is most common in a performance-rating scenario to identify the strengths and weaknesses of the ratees, they used identification goal as a control group to compare the effects of other goals.

Their result suggested that in comparison to identification goal, pursuing harmony goal increased mean ratings and decreased discriminability. In addition, the fairness goal produced higher mean ratings, and interacted with time points to influence discriminability such that discriminability decreased mid-semester but not at the end of the semester. Finally, Wong and Kwong did not observe any significant effects in the motivating goal condition.

Wang, Wong, and Kwong (2010) conducted two studies to investigate further how rater goals influence performance-ratings. Specifically, they assessed how rater goals interacted with ratee performance levels in producing performance ratings. Whereas Wong and Kwong (2007) observed the change in discriminability and mean ratings, they did not observe the patterns in which these rating changes were obtained. Wang et al. (2010) addressed this issue by observing how the rater goals affected the ratings of high performers and low performers.

In their first study, Wang et al. (2010) used a student peer review context with within-rater manipulations to examine the effects of the same four goals (identification, harmony, motivating, and fairness) used in Wong and Kwong (2007). The identification goal was used as the control condition. They found rating elevation in all experimental

conditions: the harmony, fairness, and motivating goal, compared to the identification goal. More importantly, Wang et al. found that raters elevated the scores more for low performers than medium and high performers in these conditions.

In study 2, Wang et al. (2010) had student participants rate the performance of ratees based on a 15-minute video clip using same criteria as Study 1. They found that raters deflated the ratings of high performers in fairness goal, and the raters elevated the ratings for low performers in motivating goal setting. The harmony goal manipulation had no significant effects in Study 2.

Wong and Kwong (2007) and Wang et al. (2010) both answered interesting questions regarding how rating biases occur as a function of different rater goals, however, they did not examine how these goals affect objective rating accuracy. Although it makes sense theoretically that the identification goal used as control group in both studies would result in more accurate ratings, it is still not the best measure to assess objective accuracy. Moreover, because the identification goal itself is the control group in these studies, it is impossible to assess how accurate the identification goal is compared to true ratings scores.

The current study extended the work by Murphy et al. (2004), Wong and Kwong (2007), and Wang et al. (2010) and assessed how the rater goals affect rating accuracy. Specifically, I examined the effects of four rater goals that Murphy et al. (2004) identified: focusing on identifying the strength of the ratee (strength goal), identifying the weakness of the ratee (weakness goal), being fair to the ratee (fairness goal), and

motivating the ratee (motivating goal). The harmony goal used in Wong and Kwong (2007) and Wang et al. (2010) will not be adopted for current study because it is not suitable for the non-peer rating setting of the current study.

Cognitive Ability

A meta-analysis by Schmidt and Hunter (1998) identified cognitive ability as one of the best predictors of job performance, with an average correlation of .51. The correlation is even stronger for complex jobs with correlations reaching .58 for professional-managerial jobs and .56 for high-level complex technical jobs. Whereas cognitive ability does not predict job performance on low complexity jobs nearly as well, the correlations are still moderately strong with a correlation of.40 for semi-skilled jobs and .23 for completely unskilled jobs (Schmidt & Hunter, 1998).

Because cognitive ability is a robust predictor of job performance in variety of jobs, theoretically, cognitive ability should predict performance of a cognitive task such as performance ratings. However, research findings regarding relationships between cognitive ability and accuracy of performance ratings are inconsistent. For instance, Smither and Reilley (1987) observed a curvilinear relationship between rater intelligence and accuracy of the performance ratings, such that the most intelligent raters were less accurate in their ratings than the moderately intelligent raters, and moderately intelligent raters were more accurate than the least intelligent raters. Smither and Reilley (1987) speculated that this curvilinear relationship may have been due to most intelligent raters being bored with the task, but they were unable to provide empirical support. Borman

(1979) examined various personal attributes and their influence on rating accuracy and found that verbal reasoning was the only variable that was related to both of the two job types (recruiting interviewer and manager) that he examined. Finally, Hauenstein and Alexander (1991) observed a positive relationship between cognitive ability and rating accuracy in their study.

Although the past findings have not been consistent, I suspect that there is a positive relationship between rater cognitive ability and rating accuracy. Given the fact that cognitive ability is a powerful predictor of variety of cognitive tasks, having higher cognitive ability should be advantageous in a performance rating task. In the current study, I will examine the relationship between rater cognitive ability and rating accuracy.

Personality Traits

The use of personality traits in predicting job related outcomes has been unsuccessful prior to the establishment big five personality traits (Barrick & Mount, 1991). However, the meta-analysis by Barrick and Mount demonstrated that the big five personality traits predict job performance, and many subsequent organizational researchers have incorporated personality measures in order to assess individual differences (Barrick & Mount, 1991; Bernadin, Coooke, & Villanova, 2000; Tziner, Murphy, & Cleveland, 2002; Yun, Donahue, Dudley, & McFarland, 2005).

As with other organizational studies, many of recent performance rating studies have investigated the relationships between personality attributes of the raters and their ratings. For example, Bernadin et al. (2000) observed the relationship between rater

conscientiousness, agreeableness, and rating elevation. They found that rater conscientiousness was negatively, and rater agreeableness was positively related to rating elevation. Additionally, they found that low rater conscientiousness and high rater agreeableness resulted in the highest rating elevation.

Several other researchers have tried to address the issue of rater personality traits and rating elevation. Tziner et al. (2002) examined whether rater conscientiousness moderated the effect of rater attitudes on rating behavior. They administered several rater attitudes measures, which included rating-self-efficacy, perceptions of how the performance evaluations are used, confidence in the performance evaluation system, comfort level with the performance evaluation system. Tziner and colleagues hypothesized and found that rater attitudes have significant effects on performance ratings but that these relationships are moderated by rater conscientiousness. Specifically, relationships between rater attitudes and the ratings were especially strong for raters low in conscientiousness. Conversely, raters high in conscientious tended to give ratings that were not as strongly influenced by various rater attitudes.

Yun et al. (2005) also investigated how rater personality interacts with other factors in performance evaluation setting. Their research examined rater personality (conscientiousness and agreeableness), rating formats (graphing rating scale and behavioral checklist), and rating social contexts (existence of face-to-face feedback) to determine whether these factors contributed to rating elevation. Yun and colleagues' study demonstrated a few important relationships involving and other factors. First, Yun

et al. found that raters who are high on agreeableness tended to provide elevated ratings when there was face-to-face feedback. This makes sense because agreeable people by their nature would try to avoid conflict with others. Next, Yun et al. found that use of behavioral checklist versus graphing rating scale moderated the rating elevation due to high rater agreeableness, such that raters high in agreeableness tended to elevate ratings less when using behavioral checklist than when using graphic rating scales. Finally, Yun et al. found that whereas highly agreeable raters elevated their ratings in general, such raters still distinguished between very low performing ratees. Yun et al. explained that even highly agreeable raters may have had trouble justifying rating lowest performing ratees as similar to medium and high performing ratees. Highly agreeable raters tended to rate medium and high performing ratees similarly when using graphing ratings scales and face-to-face evaluations.

The findings of these studies examining relationships between personality and performance ratings have demonstrated that conscientiousness has a negative and agreeableness a positive relationship with rating elevation (Bernadin et al., 2000; Yun et al., 2005). These studies relied on the assumption that rating elevation result in less accurate ratings. This implies that high conscientiousness results in high accuracy and that high agreeableness results in lower accuracy. However, these did not provide objective evidence to support such claims. In the current study, I will examine how these variables influence rating accuracy.

The Current Study

There were two main goals in this study. The first was to replicate the findings of the previous studies and identify which rater characteristics result in rating elevation. The other was to examine how these rater characteristics affect the rating accuracy. Most of past research on rater characteristics has focused on whether various rater characteristics contribute to rating elevation, without necessarily linking them to objective performance accuracy. Perhaps one exception is Spence and Keeping (2010). Spence and Keeping (2010) have incorporated a policy capturing approach to assess how organizational/situational characteristics and rater attributes biases the ratings. In a policy capturing study, raters examine and rate various scenarios that vary in situational and performance cues. Based on the ratings for each scenario, the researchers are able to assess how the various components in the scenarios as well as the rater attributes contribute to the observed ratings. However, although Spence and Keeping (2010) have used objective performance cues to be evaluated in their study, their study utilized straightforward description of the performance level in the scenarios (e.g. "James is an above average performer."). While this approach may be appropriate to assess how various factors bias the ratings, it is not ideal to assess how various factors affect the ability of the raters to provide accurate ratings. The current study addressed this issue by providing videos depicting varying levels of performance that the raters evaluated, rather providing explicit levels of performance in text. This allowed me to assess how the raters were able to make accurate judgments about ratees' performance based behavioral cues.

Method

Participants

A sample of 233 (31% male; 69% female) students from a Midwestern university served as supervisor raters in our study. The average age of the participants was 20.02 years (SD = 4.11). Of the participants, 52% were currently employed and 38% had experience rating employees. The average amount of months worked was 39.88 (SD = 52.63). The majority of participants were either Caucasian (60%) or African American (26%). The remaining participants were either Asian (3%), Multi-racial (5%), Hispanic (2%), or Other (4%). Only 1.7% of the data contained missing data so I decided to remove these cases from further analyses. Our final sample contained 229 raters and 2,744 observations.

Design

The current employed 3 x 4 within subject design. Participants viewed 12 unique video clips and four duplicate video clips of a waitress representing three different performance levels (low, medium, and high) in four job performance dimensions (job knowledge, cooperation, stress management, and work habits).

Measures

Rater goals. Rater goals were assessed using a four-item, five point graphic rating scale. Each item represented one of the four rater goals; motivating the ratee, identifying strength of the ratee, identifying weakness of the ratee, and being fair to the ratee (Appendix A.).

Cognitive ability test. I used the Shipley Institute of Living Scale (SILS) to assess the cognitive ability of the participants. The SILS consists of two sections: a 40-item verbal section designed to assess crystallized intelligence and a 20-item abstract reasoning section to assess fluid intelligence. In the verbal section, the participants were asked to identify synonyms using multiple choice format. In the abstract reasoning section, the participants were asked to complete the pattern by filling in the blanks with numbers or letters. Participants were given 10 minutes to complete each section. Example items from each section are shown below in Figures 1 and 2. The scores on the abstract reasoning section were doubled and added to the scores on verbal section to produce total test scores.

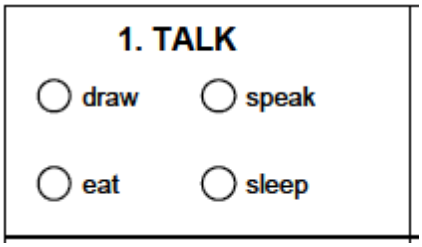

Figure 1. An example verbal item from Shipley Institute of Living Scale.

Figure 2. An example abstract reasoning item from Shipley Institute of Living Scale.

Conscientiousness. The conscientiousness measure for the current study consisted of 10 items adopted from International Personality Item Pool (http://ipip.ori.org/). Each contained a description of personality characteristics, and the participants rated the extent to which they agree or disagree with each item description

using a five point graphic rating scale. Items 6 through 10 were reverse scored (See Appendix B).

Agreeableness. The agreeableness measure for the current study consisted of 10 items adopted from International Personality Item Pool (http://ipip.ori.org/). Each item contained a description of personality characteristics, and the participants rated the extent to which they agree or disagree with each item description using a five point graphic rating scale. Items 6 through 10 were reverse scored (See Appendix C).

Demographic questionnaire. The participants' demographic information was collected using a questionnaire designed for the study. The items included questions regarding participants' age, gender, race, GPA, employment status, and employment history. For the complete list of items, see Appendix D.

Task

Participants rated a waitress' performance as portrayed in a series of video clips. The participants provided a dimension specific rating for each of the 16 video clips using 5-point graphic rating scales (See Appendix E).

Video

The video clips that I used in the current study were adopted from Lewis (2006). The videos were originally produced by Barnes-Farrell (1984) for a performance appraisal research, and were replicated by Lewis for his study. The videos were approximately 30 to 80 seconds in length, and they depicted behavioral anchors that demonstrated three levels of job performance (high, medium, and low) in several

dimensions. Four job performance dimensions (cooperation, job knowledge, maintaining performance, and work habits) were adopted for the current study.

Although the videos were designed to represent specific level of performance, a pilot study (n=32) was conducted to ensure that the videos used for the current study appropriately display the intended level of performance. Two different versions of the videos were available to represent the same level of performance in the same dimension. This allowed me to choose the best version of the video clip for each unique condition based on the pilot study ratings (See Table 1).

Procedure

At the beginning of the study, the participants completed informed consent forms (Appendix F). Then the participants were told to imagine themselves as restaurant managers who provide feedbacks to their employees. The participants then watched a series of video clips that portrayed the work performances of a waitress and provided ratings on her performance. A pause was provided after every scene to enable participants to provide ratings. After providing ratings, participants completed the SILS, followed by personality inventories, demographic questionnaires, and rater goals measures. The participants were debriefed at the end of the study.

Results

The descriptive statistics and the correlations of the variables used in the analyses are reported in Table 2. The reliability of the videos was assessed by duplicate video clips, which also functioned as calibration items. Only four duplicate video clips were

used in each session, but the duplicate video clips were alternated between sessions so that the reliability can be obtained for all 12 video clips. The *n* for the duplicate video ratings ranged from 33 to 127. The reliabilities of the video clips ranged from $r = .37$ to $r = .58$, with the mean of $r = .56$ ($SD = .11$)[1].

In order to evaluate the effects of rater characteristics on ratings, I tested several multilevel random coefficient models. Multilevel random coefficient model is an appropriate method to analyze the multilevel structure of a performance evaluation data, where the ratings are nested within raters (LaHuis & Avis 2007). First, I tested a random slope model with no cross level interaction to test whether certain rater characteristics were associated with rating elevation. The level 1 equation for this model is

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X1_{ij}) + \beta_{2j}(X2_{ij}) + R_{ij} \quad (1)$$

where $X1_{ij}$ represents the dummy coded variable of medium performance cues, and $X2_{ij}$ represents the dummy coded variable of high performance cues. The low performance cues were used as the reference group. The equation 2 through 4 completes all the components for this model.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Z1_j) + \gamma_{02}(Z2_j) \ldots + \gamma_{07}(Z7_j) + U_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + U_{1j} \quad (3)$$

$$\beta_{2j} = \gamma_{20} + U_{2j} \quad (4)$$

[1] * Although Karren and Barringer (2002) recommend using duplicate items both as calibration items and reliability check items, the use of this strategy in the current study may have led to lower reliability estimates than expected.

The variables Z1 though Z7 represent all the rater-level characteristics examined in the study. A significant positive γ coefficient for predicting intercept would indicate that the variable associated with the coefficient is associated with rating elevation, whereas a significant negative relationship would indicate a negative relationship with rating elevation. The presence of U_{1j} and U_{2j} terms show that there are random effects associated with the slope of the objective cues. The result of this analysis is presented in Table 3. The result indicated that the Strength goal was positively related to rating elevation (γ_{01} = 0.08, $t(218) = 2.70$, $p<.05$) and conscientiousness was negatively related to rating elevation (γ_{05} = -0. 01, $t(218) = -2.11$, $p<.05$). The model explained roughly 66% of the variance in performance ratings.

Next, I tested a series of models that included a cross-level interaction between the objective performance cues and each of the rater characteristics. In order to avoid excessive multicolinearity, I used a piecemeal approach, where each cross-level interactions was tested one at a time. For instance, one model would test the cross-level interaction of level 2 variable Z1, then another model would test another level 2 variable Z2, and so on. The level 1 equation and the level 2 equation for the intercept of these models are identical to those from the previous section. However, the level 2 equations for β_{1j} and β_{2j} are slightly different. The equations 5 and 6 highlight these differences.

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Z1_j) + U_{1j} \quad (5)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(Z1_j) + U_{2j} \quad (6)$$

The coefficients γ_{11} and γ_{21} represent the cross-level interaction coefficients. A significant positive γ_{11} or γ_{21} coefficient indicates that the relationship between X1 or X2 variable is stronger. Because the X1 variable was a dummy coded variable for medium performance cues and the X2 variable was a dummy coded variable for high performance cues, a significant γ_{11} or γ_{21} coefficients would mean that raters were able to better distinguish among medium performance from low performance and high performance from low performance, respectively.

Table 4 shows the coefficients for each of the models. The results indicated that strength goal was associated with higher accuracy (γ_{21} = .18, t(218) = 2.59, p<.05), as well as conscientiousness (γ_{25} = .02, t(218) = 2.48, p<.05) , agreeableness (γ_{21} = .02, t(218) = .01, p<.05), and cognitive ability (γ_{17} = .01, t(218) = 2.44, p<.05 ; γ_{27} = .01, t(218) = 2.35, p<.05). Figures 3 through 6 show the visual representations of these relationships.

Discussion

The current study examined how rater characteristics are related to rating elevations and rating accuracy. Rating elevation was defined in this context as higher mean performance ratings. Rating accuracy on the other hand, was defined as stronger discrimination among low performers, medium performers, and high performers.

The results showed that strength goal was associated with elevated ratings, whereas rater conscientiousness was associated with deflated ratings. Agreeableness was not associated with rating elevation unlike in previous studies (e.g. Bernadin et al., 2000; Yun et al., 2005). Regarding the effect of rater characteristics on accuracy, strength goal,

conscientiousness, agreeableness, and cognitive ability were all related to higher accuracy. These findings seem to show that rating levels are rather independent from rating accuracy, in a sense that the some elevated ratings could distinguish between high and low performances. This is particularly noteworthy, because previous research has somewhat equated rating elevation with rating inaccuracy. However, equating rating elevation with inaccuracy is essentially confounding rater agreement (mean change) with rater reliability (consistency). This interpretation may lead to erroneous decision in some cases. For example, raters can consistently provide higher ratings to all the ratees that can still distinguish among the different performance levels of the ratees, as observed in the current study.

In some contexts, it is not crucial that raters provide accurate rating levels as long as the ratings are consistent. For example, the goal of criterion-related validity studies is to have a reliable measure of job performance that can be predicted. The accuracy of the level of ratings is not as important as the consistency of the ratings. In this context, instructing the raters to follow a particular goal, particularly strength goal, may make the ratings more consistent. In contrast, ratings used for developmental purposes may require accurate ratings in the sense of agreement between the ratings and actual performance in order to track the improvement of individual employees. In this context, it may be necessary to clarify the performance expectations at each rating level to increase the agreement of the ratings.

It is important for organizational researchers and practitioners to be cautious when evaluating the validity of performance ratings. It is not advisable to disregard the performance ratings solely based on the fact that the ratings are elevated nor is it advisable to believe that lower mean ratings are more accurate. Instead, researchers should closely examine convergent/discriminant validity, as well as various forms of reliability to make more informed decisions. In addition, researchers may benefit from clarifying the primary use of the performance ratings, to see if rating consistency is sufficient for the purpose or if rating agreement as well as consistency, is necessary.

Another interesting finding of this study is that strength goal and weakness goal appears to be separate constructs, and that they possess different properties. Although Murphy et al. (2004) identified four factors in rater goal paradigm (identifying strengths, identifying weaknesses, being fair, and motivating), some researchers have combined the strength goal and weakness goal as identification goal in some of the previous studies (e.g. Wong & Kwong, 2007; Wang et. al., 2010). However, the current study only found a moderate correlation between strength goal and weakness goal. Moreover, strength goal was associated with rating elevation and rating accuracy while weakness goal is not associated with either. Based on these findings, I encourage that researchers distinguish between strength goal and weakness goal, rather than combining them into one construct.

Limitations and Future Research

The current study employed student sample, many of whom had limited work experiences. Although use of student sample is common in organizational research, this

could have influenced the results of the current findings. It may be beneficial to replicate the study using professional sample to see if the findings are reliable.

In addition, the use of hypothetical ratees rather than employing real people may have had some influence on my findings. The use of policy capturing type scenarios in the current study allowed me to evaluate the observed ratings against the objective performance levels. However, although I adopted a series of videos to enhance the realism, it was still apparent to the raters that the scenarios were hypothetical. This may have reduced the rating elevations typical of many performance evaluations, which could explain why agreeableness did not result in rating elevation in the current study unlike previous studies by Bernadin et al. (2000) and Yun et al. (2005).

Another potential limitation of the current study is that the rater goals were measured using four-item questionnaire in order to reduce the survey length. Each of the four items in the questionnaire corresponded to one of the four rater goals. This may be of concern to some researchers, since only one item was used for each of the four rater goals. However, given the clarity of the rater goals assessed in the current study, the use of one item measure was likely not a problem. Use of single item measures may be unacceptable for a complex multi-facet construct but may be acceptable for homogeneous constructs (Loo, 2002). In addition, many studies have shown that use of a single item measures can be valid in assessing various psychological constructs such as job satisfaction, brand attitudes, and global self-esteem (e.g. Bergkvist & Rossiter, 2007;

Dolbier, Webster, McCalister, Mallown, & Steinhardt, 2005; Robins, Hendin, & Trzesniewski, 2001, Wanous, Reichers, & Hudy, 1997).

An additional topic that future research could investigate is the consistency of rater goals within a performance evaluation process and across time. Currently, the norm in rater goal research is based on the assumption that the rater goal is somewhat stable across ratees. However, it is quite possible that a rater may adopt different goals for different ratees, even within the same performance evaluation process. In addition, it may be interesting to see if rater goal is stable over time, or if it is variable. Examining these characteristics regarding rater goal could expand the understanding of the topic and possibly improve the utility of the construct.

Conclusion

The current study examined how the rater-level characteristics affect the performance evaluation ratings. The results showed that focusing on the strength of the ratee was associated with elevated ratings, while rater conscientiousness was associated with deflated ratings. However, both of those rater characteristics, along with rater agreeableness and rater cognitive ability, were associated with higher rating accuracy.

References

Barnes-Farrell, J. L. (1984). The development of a laboratory measure of accuracy in performance appraisal. *Personnel Psychology, 38,* 335-345.

Barrick, M., & Mount, M. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

Barrick, M., Mount, M., & Judge, T. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 9–30.

Bernardin, J., Cooke, D., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology, 85,* 232–236.

Bretz, R., Milkovich, G., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18,* 321–352.

Borman, W. C. (1979). Individual differences correlates of accuracy in evaluating others' performance effectiveness. *Applied Psychological Measurement, 3,* 103-115.

Cardy, R., & Dobbins, G. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology, 71,* 672–678.

Dolbier, C.L., Webster, J.L., McCalister, K. T., Mallown, M. W., & Steinhardt, M. A. (2005). Reliability and validity of a single-item measure of job satisfaction. *American Journal of Health Promotion, 19,* 194-198.

Hauenstein, N. M. A. (1992). An information processing approach to leniency in performance judgments. *Journal of Applied Psychology*, *77*, 485–493.

International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (http://ipip.ori.org/). Internet Web Site.

Kane, J., Bernardin, J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal, 38,* 1036–1051.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of the policy-capturing methodology in organizational research: Guidelines for research and practice. *Organizational Research Methods, 5,* 337-361.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, *70*, 56–65.

La Huis, D. M., & Avis, J. M. (2007). Using multilevel random coefficient modeling to investigate rater effects in performance ratings. *Organizational Research Methods, 10,* 97-107.

Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology, 17,* pp. 68-75.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44,* pp. 175-184.

Lewis, W. R.(2006). *Effect of mode of observation on accuracy of performance appraisal judgments* (Unpublished master's thesis). University of Connecticut, CT.

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology, 89,* 158–164.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Global self-esteem: Construct validation of a single-item measure and the rosenberg self-esteem scale. *Personality and Social Psychology Bulletin, 27,* 151-161.

Schmidt, F. E., & Hunter, J. E. (1998). The validity and utility of selection practices in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Sinclair, R. (1988). Mood, categorization, breadth, and performance appraisal: The effects of order of information acquisition and affective state on halo, accuracy, information, retrieval, and evaluations. *Organizational Behavior and Human Decision Processes, 42,* 22–46.

Smither, J. W., & Reilly, R. R. 1987. True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes, 40,* 369-391.

Spence, J. R., & Keeping, L. M. (2010). The impact of non-performance information on ratings of job performance: A policy-capturing approach. *Journal of rganizational Behavior, 31,* 587-608.

Tziner, A., Murphy, K., & Cleveland, J. (2002). Does conscientiousness moderate the relationship between attitudes and beliefs regarding performance appraisal and rating behavior? *International Journal of Selection and Assessment, 10,* 218–224.

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

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction, How good are single-item measures? *Journal of Applied Psychology, 82,* 247-252.

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goal on rating patterns: Evidencefrom an experimental field study. *Journal of Applied Psychology, 90,* 284-294.

Yun, G., Donahue, L., Dudley, N., & McFarland, L. (2005). Rater personality, rating format, and social context: Implications for performance appraisal ratings. *International Journal of Selection and Assessment, 13,* 97–107.

Table 1.

Pilot Means (and Standard Deviations)

	Dimension (1-5)	Overall (1-5)	Difficulty Rating (1-3)
CL1	1.22 (.42)	1.56 (.95)	1.19 (.54)
CL2	1.47 (.84)	1.81 (1.12)	1.16 (.52)
CM1	4.66 (.65)	4.47 (.72)	1.13 (.42)
CM2	4.25 (.84)	3.75 (1.08)	1.41 (.67)
CH1	4.53 (.67)	4.00 (.95)	1.19 (.54)
CH2	4.78 (.49)	4.56 (.80)	1.13 (.43)
MPL1	2.13 (.94)	2.25 (1.14)	1.28 (.58)
MPL2	2.09 (.73)	2.50 (.76)	1.34 (.55)
MPM1	3.28 (1.09)	3.38 (1.07)	1.72 (.81)
MPM2	2.94 (.88)	3.06 (.88)	2.09 (.73)
MPH1	4.66 (.65)	4.44 (.84)	1.28 (.52)
MPH2	4.34 (.97)	4.31 (1.03)	1.31 (.64)
WHL1	2.28 (1.20)	2.13 (1.07)	1.53 (.76)
WHL2	1.34 (.79)	1.65 (.92)	1.28 (.63)
WHM1	3.59 (1.27)	3.41 (1.16)	2.03 (.82)
WHM2	2.06 (.84)	2.72 (1.05)	1.28 (.58)
WHH1	3.97 (1.23)	4.53 (.76)	1.50 (.67)
WHH2	2.88 (1.36)	2.97 (1.17)	1.59 (.67)
JKL1	1.19 (.54)	1.44 (.72)	1.06 (.35)
JKL2	1.50 (.62)	1.59 (.98)	1.09 (.39)
JKM1	2.38 (.79)	2.84 (.95)	1.28 (.52)
JKM2	3.66 (1.26)	2.28 (1.40)	1.34 (.55)
JKH1	4.69 (.69)	4.50 (.67)	1.13 (.42)
JKH2	4.66 (.65)	4.31 (.82)	1.16 (.45)

Note. n = 32. C = Cooperation, MP = Maintaining Performance, WH = Work Habits, JK = Job Knowledge. L = Low performance, M = Medium performance, H = High performance.

Table 2.

Means, Standard Deviations, and Correlations of the Variables.

Variable	Mean	SD	1	2	3	4	5	6	7	8
Level 1										
Ratings	3.10	1.43	___							
Level 2										
Strength Goal	4.01	0.76	0.03	___						
Weakness Goal	3.87	0.88	0.00	0.29	___					
Fairness Goal	4.10	0.94	0.02	0.30	0.23	___				
Motivating Goal	3.49	0.91	0.00	0.31	0.14	0.25	___			
Conscientiousness	37.39	5.64	-0.03	0.14	0.07	-0.01	0.15	**0.88**		
Agreeableness	40.38	5.72	0.01	0.22	0.02	0.11	0.15	0.17	**0.82**	
Cognitiveability	55.56	9.10	0.00	0.08	0.14	0.13	0.04	-0.06	0.08	**0.86**

Note. Cronbach's α coefficients appear on the diagonal in boldface.
SD, standard deviation.
n = 227 participants, *N*= 2720 observations.

Table 3.

Random Slope Model

Variable	β_u [a]	SE [b]	*t*	Variance Component
Intercept, γ_{00}	1.72	0.24	7.29*	0.15*
Level 1				
Medium Performance, β_1	1.29	0.04	29.98*	0.12*
High Performance, β_2	2.85	0.05	52.44*	0.37*
Level 2				
Strength Goal, γ_{01}	0.08	0.03	2.70*	
Weak Goal, γ_{02}	-0.01	0.02	-0.49	
Fairness Goal, γ_{03}	0.03	0.02	1.10	
Motivating Goal, γ_{04}	-0.02	0.02	-0.97	
Conscientiousness, γ_{05}	-0.01	0.00	-2.11*	
Agreeableness, γ_{06}	0.00	0.00	0.69	
Cognitive Ability, γ_{07}	0.00	0.00	-0.94	

Note. N=2708, n=226.

[a] Unstandardized coefficients.

[b] Standard Error.

*p<.05.

Table 4.

Random Slope Models with Cross-Level Interactions

Cross-Level Interactions	β_u [a]	SE [b]	*t*
Low vs. Med x Strength Goal, γ_{11}	0.11	0.06	1.90
Low vs. High x Strength Goal, γ_{21}	0.18	0.07	2.59*
Low vs. Med x Weakness Goal, γ_{12}	-0.03	0.05	-0.70
Low vs. High x Weakness Goal, γ_{22}	-0.02	0.06	-0.26
Low vs. Med x Fairness Goal, γ_{13}	0.02	0.05	0.36
Low vs. High x Fairness Goal, γ_{23}	0.10	0.06	1.75
Low vs. Med x Motivating Goal, γ_{14}	0.03	0.05	0.65
Low vs. High x Motivating Goal, γ_{24}	0.06	0.06	1.03
Low vs. Med x Conscientiousness, γ_{15}	0.01	0.01	1.76
Low vs. High x Conscientiousness, γ_{25}	0.02	0.01	2.48*
Low vs. Med x Agreeableness, γ_{16}	0.01	0.01	1.60
Low vs. High x Agreeableness, γ_{26}	0.02	0.01	2.01*
Low vs. Med x Cognitive Ability, γ_{17}	0.01	0.01	2.44*
Low vs. High x Cognitive Ability, γ_{27}	0.01	0.01	2.35*

Note. N=2708, n=226.

[a] Unstandardized coefficients.

[b] Standard Error.

*p<.05.

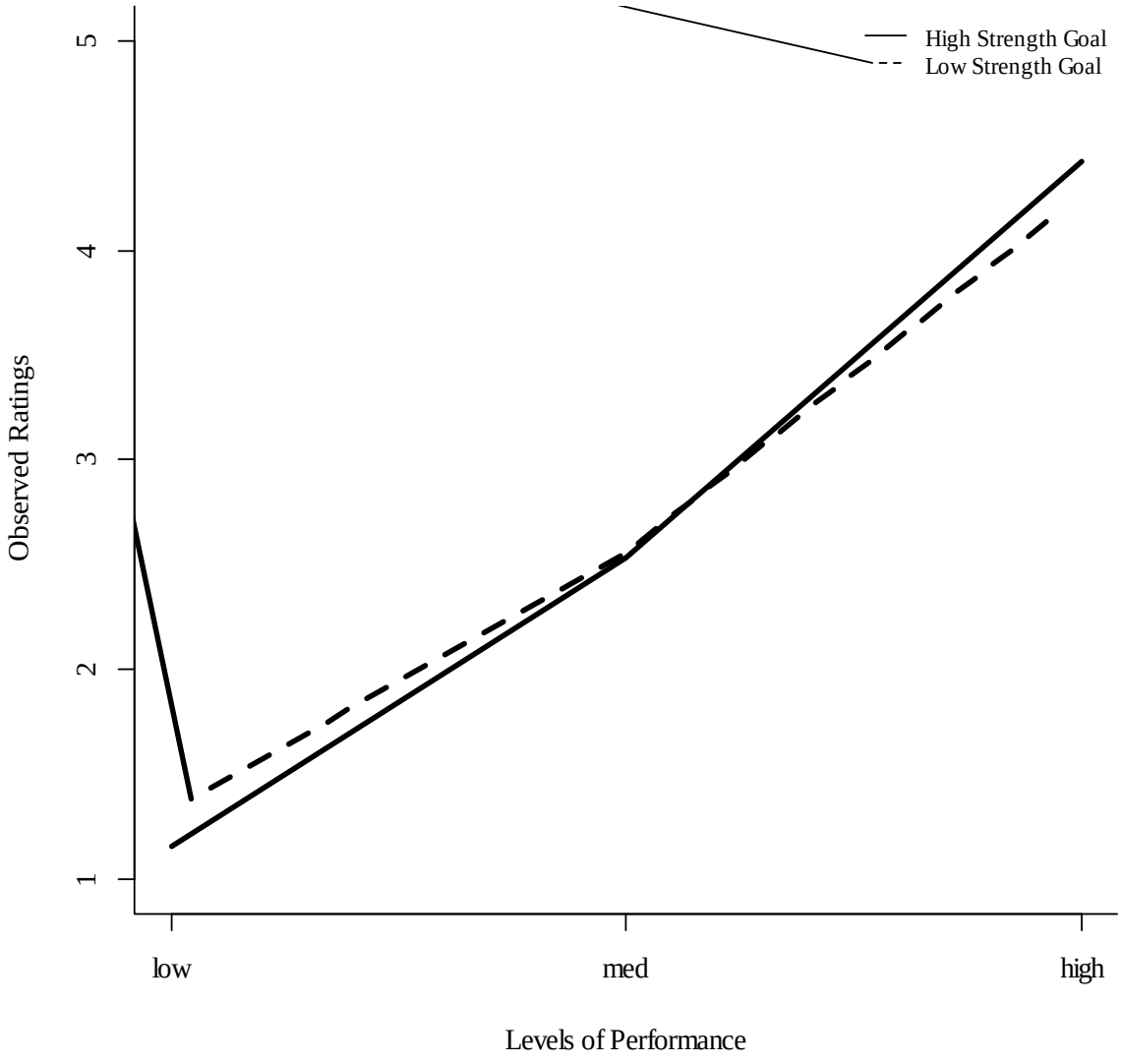

Figure 3. Cross level interaction between objective performance cues and levels of strength goal. The solid line represents one standard deviation above the mean while the dotted line represents one standard deviation below the mean.

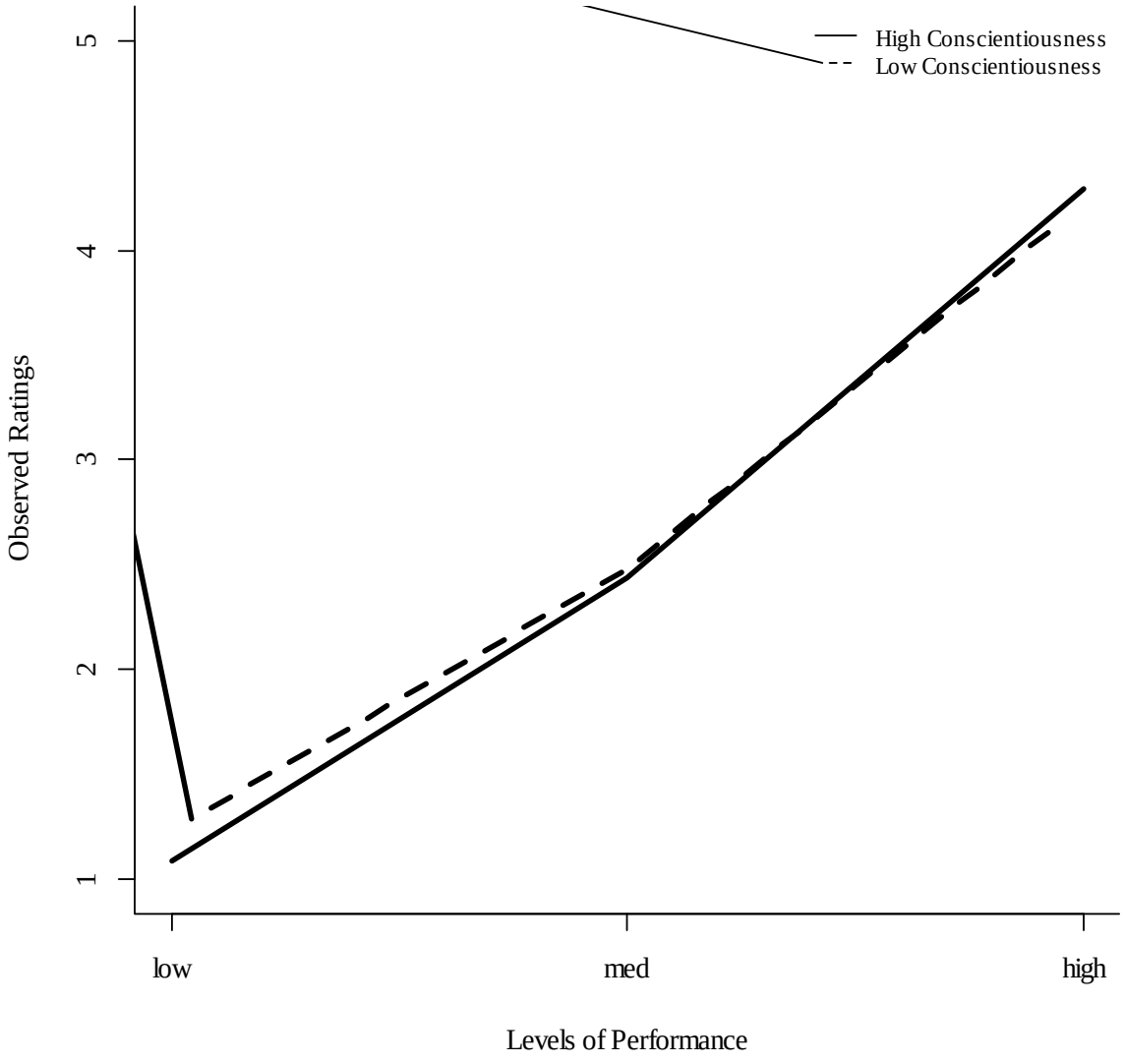

Figure 4. Cross level interaction between objective performance cues and levels of conscientiousness. The solid line represents one standard deviation above the mean while the dotted line represents one standard deviation below the mean.

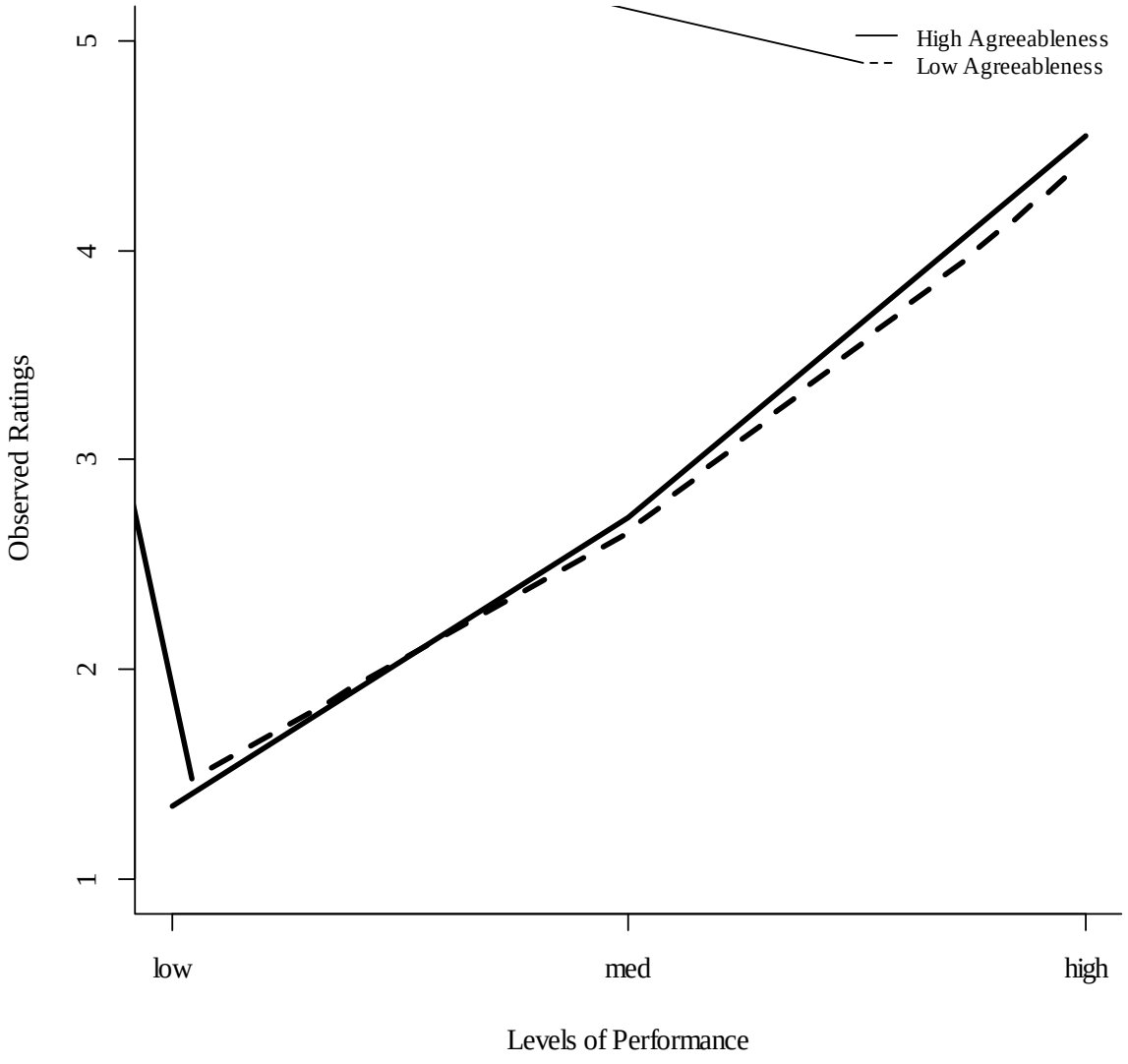

Figure 5. Cross level interaction between objective performance cues and levels of agreeableness. The solid line represents one standard deviation above the mean while the dotted line represents one standard deviation below the mean.

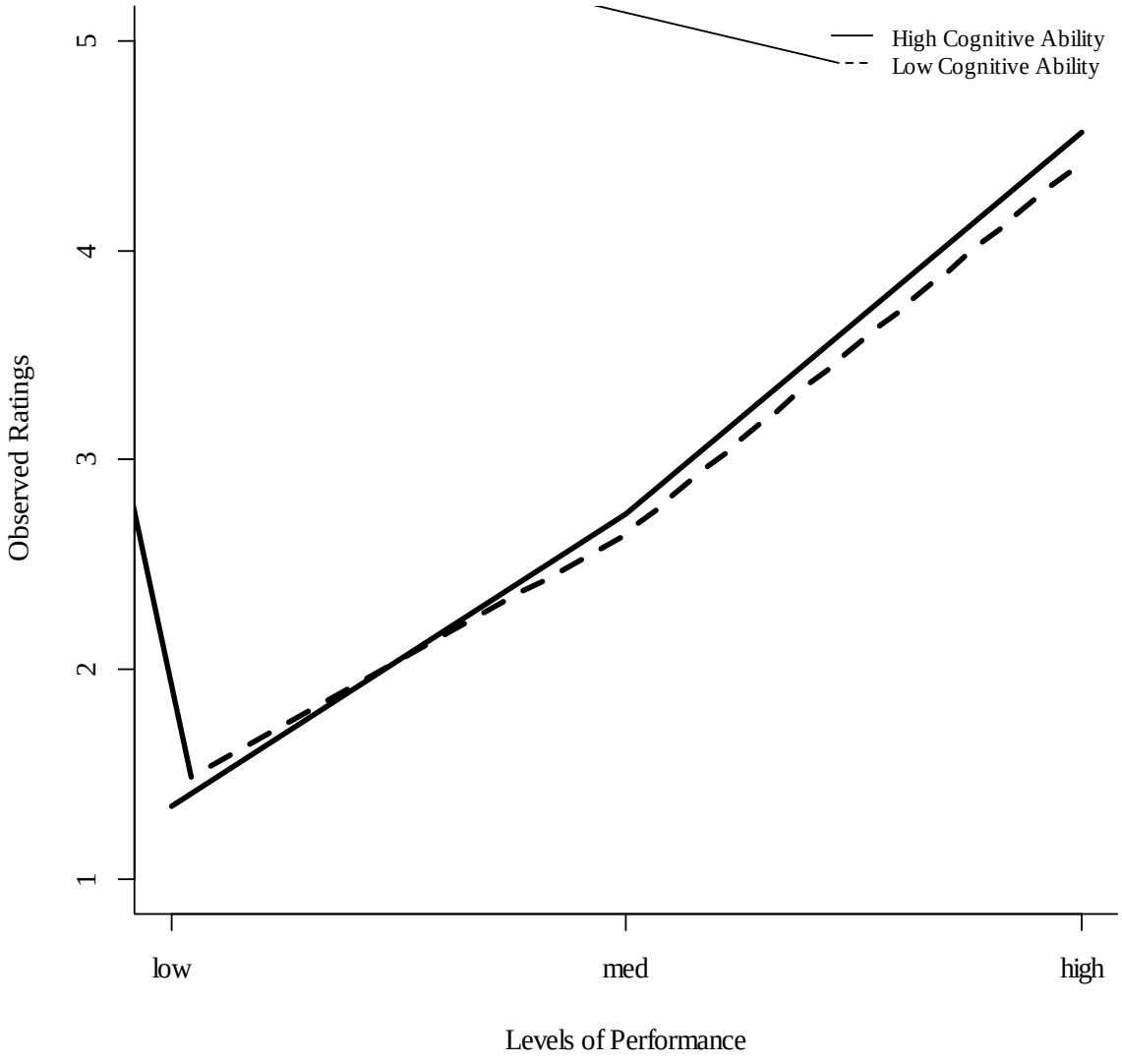

Figure 6. Cross level interaction between objective performance cues and levels of cognitive ability. The solid line represents one standard deviation above the mean while the dotted line represents one standard deviation below the mean.

Appendix A

Please answer honestly and as accurately as possible.

Rater Goals	**Ratings**				
While rating the server, I...	**strongly disagree**	**disagree**	**neutral**	**agree**	**strongly Agree**
...focused on keeping the server motivated for the job.	**1**	**2**	**3**	**4**	**5**
...focused on identifying the server's strengths.	**1**	**2**	**3**	**4**	**5**
...focused identifying the server's weaknesses.	**1**	**2**	**3**	**4**	**5**
...focused on being fair and unbiased.	**1**	**2**	**3**	**4**	**5**

Appendix B

Please circle the number that matches your level of agreement with each statement.

Personality Test	**Ratings**				
	strongly disagree	**disagree**	**neutral**	**agree**	**strongly Agree**
1. I am always prepared.	**1**	**2**	**3**	**4**	**5**
2. I pay attention to details.	**1**	**2**	**3**	**4**	**5**
3. I get chores done right away.	**1**	**2**	**3**	**4**	**5**
4. I like order.	**1**	**2**	**3**	**4**	**5**
5. I follow a schedule.	**1**	**2**	**3**	**4**	**5**
6. I am exacting in my work.	**1**	**2**	**3**	**4**	**5**
7. I leave my belongings around.	**1**	**2**	**3**	**4**	**5**
8. I make a mess of things.	**1**	**2**	**3**	**4**	**5**
9. I often forget to put things back in their proper place.	**1**	**2**	**3**	**4**	**5**
10. I ignore or do not do my duties.	**1**	**2**	**3**	**4**	**5**

Appendix C

Please circle the number that matches your level of agreement with each statement.

Personality Test	**Ratings**				
	strongly disagree	**disagree**	**neutral**	**agree**	**strongly Agree**
1. I am interested in people.	**1**	**2**	**3**	**4**	**5**
2. I sympathize with other's feelings.	**1**	**2**	**3**	**4**	**5**
3. I have a soft heart.	**1**	**2**	**3**	**4**	**5**
4. I take time out for others.	**1**	**2**	**3**	**4**	**5**
5. I feel others' emotions.	**1**	**2**	**3**	**4**	**5**
6. I make people feel at ease.	**1**	**2**	**3**	**4**	**5**
7. I am not really interested in others.	**1**	**2**	**3**	**4**	**5**
8. I insult people.	**1**	**2**	**3**	**4**	**5**
9. I am not interested in other people's problems.	**1**	**2**	**3**	**4**	**5**
10. I feel little concern for others.	**1**	**2**	**3**	**4**	**5**

Appendix D

Demographic Questionnaire

1. Age: _____

2. Gender (Please check the appropriate box)

- ☐ Male
- ☐ Female

3. Race (Please check the appropriate box)

- ☐ Pacific Islander
- ☐ African American
- ☐ Hispanic
- ☐ Native American
- ☐ Caucasian
- ☐ Asian
- ☐ Multi-racial
- ☐ Others

4. What is your GPA? _____

5. Have you ever been employed? **Circle yes or no**. (Yes/No)

If yes, how many years have you worked total? _____ years _____months.

6. Are you currently employed? **Circle yes or no**. (Yes/No)

If yes, how many hours on average do you work a week? _____ hours.

7. Have you ever given performance evaluation? **Circle yes or no**. (Yes/No)

If yes, how many years total have you worked in a position that gives performance evaluation? _____ years _____months.

8. Do you currently hold a position that provides performance evaluation to other employees? **Circle yes or no**. (Yes/No)

Appendix E

Please circle the number that reflects the performance of the server.

Performance Criteria: Scene 1	Ratings				
	Below Average		Average		Above Average
Work Habits: The server's reliability/availability was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**
Performance Criteria: Scene 2	**Ratings**				
	Below Average		**Average**		**Above Average**
Maintaining Performance The server's ability to cope with stress was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**
Performance Criteria: Scene 3	**Ratings**				
	Below Average		**Average**		**Above Average**
Job Knowledge The server's knowledge regarding menus was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**
Performance Criteria: Scene 4	**Ratings**				
	Below Average		**Average**		**Above Average**
Cooperation The server's willingness to help the coworkers was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 5	Ratings				
	Below Average		Average		Above Average
Work Habits: The server's reliability/availability was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 6	Ratings				
	Below Average		Average		Above Average
Job Knowledge The server's knowledge regarding menus was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 7	Ratings				
	Below Average		Average		Above Average
Cooperation The server's willingness to help the coworkers was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 8	Ratings				
	Below Average		Average		Above Average
Cooperation The server's willingness to help the coworkers was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 9	Ratings				
	Below Average		Average		Above Average
Job Knowledge The server's knowledge regarding menus was:	1	2	3	4	5
Overall Performance The overall effectiveness of the server was:	1	2	3	4	5

Performance Criteria: Scene 10	Ratings				
	Below Average		Average		Above Average
Cooperation The server's willingness to help the coworkers was:	1	2	3	4	5
Overall Performance The overall effectiveness of the server was:	1	2	3	4	5

Performance Criteria: Scene 11	Ratings				
	Below Average		Average		Above Average
Work Habits: The server's reliability/availability was:	1	2	3	4	5
Overall Performance The overall effectiveness of the server was:	1	2	3	4	5

Performance Criteria: Scene 12	Ratings				
	Below Average		Average		Above Average
Maintaining Performance The server's ability to cope with stress was:	1	2	3	4	5
Overall Performance The overall effectiveness of the server was:	1	2	3	4	5

Performance Criteria: Scene 13	Ratings				
	Below Average		Average		Above Average
Job Knowledge The server's knowledge regarding menus was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 14	Ratings				
	Below Average		Average		Above Average
Work Habits: The server's reliability/availability was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 15	Ratings				
	Below Average		Average		Above Average
Maintaining Performance The server's ability to cope with stress was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Performance Criteria: Scene 16	Ratings				
	Below Average		Average		Above Average
Maintaining Performance The server's ability to cope with stress was:	**1**	**2**	**3**	**4**	**5**
Overall Performance The overall effectiveness of the server was:	**1**	**2**	**3**	**4**	**5**

Appendix F

CONSENT TO PARTICIPATE IN RESEARCH

Department of Psychology, Wright State University, Dayton, OH 45435

TITLE OF THE STUDY	Accuracy of Performance Evaluations
PURPOSE OF STUDY	The purpose of this research study is to measure the factors influencing performance appraisals.
ACTIVITIES/PROCEDURES	I will be completing a short personality questionnaire. I will also be watching videos and assessing the performance of a restaurant's wait staff using paper questionnaires. Participation should last no more than 1 hour.
BENEFITS AND RISKS	There are no known risks or discomforts. There are no direct benefits for participation in this research study.
CONFIDENTIALITY	Any information about me obtained from this study will be kept strictly confidential and I will not be identified in any report or publication. No identifying information about me will be recorded on any of the questionnaires.
COMPENSATION	For my participation, I will receive, for each half-hour or part thereof, 1 research participation course credit, for a total of 2 credits for full completion of the study.
FREEDOM TO WITHDRAW	I am free to refuse to participate in this study or to withdraw at any time. My decision to participate or not to participate will not adversely affect my standing at this institution or cause a loss of benefits to which I might otherwise be entitled. There is no penalty of any kind for either non-participation or withdrawal at any time.
AVAILABILITY OF RESULTS	A summary of the results of this study may be requested by contacting the researchers listed below after March 15, 2012. The summary will show only aggregate (combined) data. No individual results will be available.
INVESTIGATOR AVAILABILITY	If I have questions about this research study, I can contact the investigators David LaHuis and Shotaro Hakoyama at 937-775-2391. If I have general questions about giving consent or my rights as a research participant in this research study, I can call the Wright State University Institutional Review Board at 937-775-4462.

CONSENT

My signature below means that I have freely agreed to participate in this investigational study.

SIGNATURE/DATE LINES

_____

(Printed Name of Participant)

_____

(Participant's Signature)
Date

Shotaro Hakoyama 937-775-2391

(Typed Name of Principal Investigator) Telephone
Date

David LaHuis 937-775-2391

(Typed Name of Faculty Advisor) Telephone
Date