RATER RATING-LEVEL BIAS AND ACCURACY IN PERFORMANCE APPRAISALS: THE IMPACT OF RATER PERSONALITY, PERFORMANCE MANAGEMENT COMPETENCE, AND RATER ACCOUNTABILITY

H. JOHN BERNARDIN, STEPHANIE THOMASON, M. RONALD BUCKLEY, AND JEFFREY S. KANE

*We studied the problem of rating-level bias and rating accuracy among retail managers of a Fortune 500 retailer. Hypotheses were tested regarding the relationship among managers' Five-Factor Model (FFM) personality characteristics, their competence in performance management, and their levels of bias and accuracy in appraisals made in situations differing on levels of rater accountability. Associate store managers (*N *= 125) rated subordinates, peers and managers under conditions of high and low rater accountability. We found support for the stability of rating-level bias across rating situations. Raters' levels of agreeableness and assertiveness were related to mean rating levels across situations, and U-shaped relationships were found in predicting one measure of rating accuracy such that high and low levels of these two traits were related to greater rating inaccuracy. Conscientiousness scores were significantly (and negatively) correlated with highly accountable mean ratings of subordinates. Performance management competence was related to rating-level bias in both high- and low-accountability conditions and contributed incremental validity in the prediction of rating level and rating accuracy. Our results indicate that the most lenient raters are more agreeable, less assertive, and less competent in performance management. These raters may also be less accurate. © 2015 Wiley Periodicals, Inc.*

Keywords: rating-level bias, leniency, accountability, stability, Five-Factor Model of personality, performance management competence

Correspondence to: H. John Bernardin, Florida Atlantic University, Department of Management Programs, College of Business, Boca Raton, Florida 33431, Phone: 561-297-3640. E-mail: bernardi@fau.edu

Human Resource Management, March–April 2016, Vol. 55, No. 2. Pp. 321–340

Published online in Wiley Online Library (wileyonlinelibrary.com).
DOI:10.1002/hrm.21678

The tendency for managers and supervisors to be lenient in their performance appraisals remains one of the most significant problems related to performance appraisal systems (Kneeland, 1929; Pulakos & O'Leary, 2011). The magnitude of rater bias effects such as leniency is well documented (Hoffman, Lance, Bynum, & Gentry, 2010; O'Neill, Goffin, & Gellatly, 2012). This form of bias can result in rating inaccuracy and undermine validity because the "true rank ordering among targets is thereby obscured" (Hoyt, 2000, p. 65). Limited empirical research on correlates of rater-level bias and accuracy has focused on the influences of particular rater attributes, especially personality, appraisal system characteristics, especially rater accountability factors, and the relationship between rater rating-level bias and rating accuracy (e.g., Bernardin, Tyler, & Villanova, 2009; Bono, Hooper, & Yoon, 2012).

The purpose of this research was to study rater-level bias and leniency in particular among managers of a Fortune 500 retailer. We addressed the following research questions: (1) What is the stability of individual rater rating-level biases across different rating situations? (2) What is the relationship between rating-level bias and rating accuracy? (3) To what extent are a rater's personal characteristics (i.e., rater personality and rater competence) related to rating-level bias and accuracy? and (4) Do performance measurement accountability factors moderate the relationships between these rater personal characteristics, rating-level bias, and rating accuracy?

Theories of Rater-Level Bias

Guilford (1954) was the first to proffer the idea that the tendency to be lenient (or harsh) on the part of raters is stable across rating situations and that a rater's personality is the underlying cause of these biases. Few studies have investigated the stability of rater rating levels across different rating situations (Jawahar, 2001).

Spence and Keeping (2011) describe leniency as "a conscious bias that is the result of a motivated behavior, such as the desire to avoid discomfort" (p. 92). This motivation, including a motivation to rate inaccurately, derives from raters' particular attributes, especially personality, and the context for appraisal, especially rater accountability. They also cite Hauenstein (1992), who depicted leniency as a "non-conscious" bias and mainly a function of raters' attention, processing, storage, and retrieval biases. Spence and Keeping (2011) use "regulatory focus theory" (RFT; Higgins, 1997) to assess motivations to rate certain ways, including leniently. According to RFT, managers are motivated to move closer to a desired reference point or move farther away from an undesirable reference point. Spence and Keeping (2011, pp. 91–92) wrote, "Viewing the rating distortion literature from the perspective of RFT helps to understand the motives underlying rater behavior as either striving towards a goal or attempting to avoid an undesired result."

Are Rater Rating-Level Effects Stable Across Situations?

Kane, Bernardin, Villanova, and Peyrefitte (1995) reported significant correlations in mean rating levels among managers rating different ratees performing the same jobs but using very different rating formats, with an average correlation of .46 across three studies. Wood, Harms, and Vazire (2010) found reliable "perceiver effects" across ratees and rated attributes in three rating studies. A "positivity factor," defined in their studies as leniency is typically defined (Bernardin, Cooke, & Villanova, 2000), transcended particular ratees, rating situations, the particular ratee characteristics being rated, and the various relationships of the perceivers to the targets. No studies have investigated rater rating-level effects across different hierarchical rating levels (Jawahar, 2001). Our first major research question then is:

> *To what extent are managers' mean rating levels correlated across different ratee hierarchical rating levels?*

The Relationship Between Mean Rating Level and Rating Accuracy

Leniency is often the focus of field studies because "true scores of performance are not available" (Ng, Koh, Ang, Kennedy, & Chan, 2011, p. 1037). Because of the measurement problems in distinguishing between rater rating-level bias and differences in "true" performance levels across rated groups or ratees (i.e., a particular manager's subordinates), an important research question is to what extent mean rater rating levels are related to measures of rating accuracy. Do more lenient (or harsh) raters, defined by the mean rating across all whom s/he rates, also rate less accurately, as defined in a number of ways, including a rater's rating deviations from the mean levels of ratings on the same people by all other raters on the same performance dimensions (Murphy & Cleveland, 1995)?

Research on this subject has suggested negligible relationships between measures of rating leniency and rating accuracy (Murphy & Cleveland, 1995; Sulsky & Balzer, 1988). However, Bernardin et al. (2009) found mean rating levels

were positively correlated with an average deviation score (compared to expert raters) and also a "correlational accuracy" score using the same expert ratings as the comparative "true score." No one has investigated these relationships in a setting outside of the college classroom. In this study, we used ratings from other sources to define "true score" representations of ratee performances for the derivation of accuracy measures (Murphy & Cleveland, 1995, p. 290). As recommended by Bernardin et al. (2009), we explored the possibility of nonlinear relationships between rating levels and particular measures of rating accuracy. Our second major research question is:

> *To what extent are managers' mean rating levels related to measures of their rating accuracy?*

Raters' Personal Characteristics, Rating-Level Bias, and Accuracy

Given some stability in a manager's tendency to be lenient and/or less accurate, the next important question is whether a manager's personal attributes (e.g., personality and competencies) are related to rating-level bias and rating accuracy. Early research indicates cognitive and experiential factors are related to accuracy (e.g., Borman, 1979). More recent research indicates raters' personalities are related to rating-level bias and accuracy and that the relationships may be moderated by the rating context, especially the role of rater accountability factors (e.g., Mero, Guidice, & Brownlee, 2007). There is little research investigating the personal characteristics of practicing managers and their rating proclivities.

Rater Personality and Mean Rating Level

Research on personality and rating bias has focused on the "Big Five" personality traits from the Five-Factor Model (FFM; Costa & McCrae, 1992). The most reliable and robust findings concern the agreeableness (A) trait of the FFM (Bernardin et al., 2000; Bernardin et al., 2009). According to Costa and McCrae (1992), more agreeable workers tend to be more empathetic to others and may be more focused on social approval in situations that require interdependence when more assertiveness may be required to solve problems (Costa & McCrae, 1992; Graziano & Eisenberg, 1997).

Bartells and Doverspike (1997) found that sensitivity and warm-heartedness, major attributes of agreeableness, were positively correlated with mean rating levels by assessors from an assessment center. Wood et al. (2010) found higher levels of raters' agreeableness were related to more positive ratings across several rating contexts. More agreeable raters were more likely to rate their peers as more agreeable and conscientious. Bono et al. (2012) found that subordinates' agreeableness was positively correlated with ratings of supervisors. We thus expect that the more agreeable managers will rate more leniently than other raters and that the less agreeable will rate more "harshly." We therefore state the following:

Hypothesis 1a: Raters' agreeableness will be positively correlated with mean rating levels of subordinates, peers and supervisors under varying conditions of accountability.

Research regarding conscientiousness (C) and rating level has revealed inconsistent results. While Bernardin et al. (2000) and Bernardin et al. (2009) found that student raters' C scores were negatively correlated with the average performance ratings these students gave their peers, Bono et al. (2012) found a positive correlation between subordinates' levels of C and their mean rating levels of their leaders' traits.[1] Because of these inconsistent results, we state no hypotheses regarding the C trait and rating outcomes.

We explored the possibility of nonlinear relationships between rating levels and particular measures of rating accuracy.

Rater Personality and Rating Accuracy

Bernardin et al. (2009) surmised that the relationship between a manager's personality and rating bias may be U-shaped. Indeed, recent research has established curvilinear relationships between certain personality traits and managerial performance levels on task performance, organizational citizenship behavior, and counterproductive behavior (Le et al., 2011). It may be that both high and low levels of certain rater traits are related to rating accuracy and that the particular relationship depends on the particular operational definition of the accuracy measure. For example, agreeableness may be the one FFM trait that has a U-shaped relationship to a measure of rating accuracy that reflects an absolute deviation from a grand mean rating level across raters on the same targets. Low levels of agreeableness may predict more severe (i.e., lower than average) ratings, and high levels of agreeableness may predict higher than average ratings. These more extreme scores in either direction may thus predict greater deviations from the "true scores" of performance on the same ratees. Thus:

Hypothesis 1b: Both high and low levels of agreeableness are related to higher levels of rating inaccuracy.

Bernardin et al. (2009) also found that "individuals who are both more agreeable and less conscientious represent the worst combination of rater traits for effective ratings. These particular raters made the most lenient and least accurate ratings" (p. 306). This finding is compatible with other research showing that interacting FFM factors may do a better job in the prediction of more narrowly defined criterion measures such as rating-level bias and rating accuracy. We also investigate whether combinations of FFM factors (namely, agreeableness and conscientiousness) plus one particular underlying FFM facet (namely, assertiveness) can provide incremental validity beyond A and C in the prediction of mean rating-level accuracy.

Rater Assertiveness, Rating Level, and Accuracy

Sackett and Lievens (2008) concluded that prediction can be improved by combinations of FFM factor scores plus selected underlying FFM facets when more theoretically compatible criteria are studied. In addition to the Big Five traits, we also focused on assertiveness, a FFM subtrait of extraversion (Costa & McCrae, 1992). Assertiveness is "a person's tendency to actively defend, pursue, and speak out for his or her own interests" (Ames & Flynn, 2007, p. 307).

While most research on extraversion has not found this trait to be correlated with rating level (the one exception is Bono et al., 2012), it has been reported that above-average (but not very high) levels of assertiveness are related to leadership effectiveness (Ames & Flynn, 2007). Assertiveness is the one underlying five-factor facet that has not been studied in the context of individual rating behavior but that has been linked to behavioral and cultural variables shown to be related to rating-level bias (Atwater, Wang, Smither, & Fleenor, 2009).

Along with the FFM domain traits, we focused on this particular facet for several reasons.[2] Assertiveness is an important and distinguishable dimension within the extraversion domain. Using exploratory factor analysis, DeYoung, Quilty, and Peterson (2007) identified two divergent extraversion dimensions, which they labeled *assertiveness* and *enthusiasm*. Enthusiasm comprised social and affective components, while assertiveness comprised items relating to social dominance and leadership ambitions. Research has also noted the importance of separating the interpersonal affiliation dimensions within the extraversion trait from the agency (assertiveness) aspects (Depue & Collins, 1999). There is also evidence that this particular facet may be related to rating bias. For example, Borman and Hallam (1991) found that the tendency to be critical, an attribute of assertiveness (Costa & McCrae, 1992), was related to lower ratings of work sample specimens. Bartells and Doverspike (1997) found that a measure of assertiveness was negatively correlated with leniency among assessors of an assessment center. Ng et al. (2011) found that more individualistic raters, an orientation correlated with assertiveness, rated less leniently in a multisource rating system. Wood et al. (2010) also found that students who self-reported as "controlling and dominant" rated less leniently.

The addition of the assertiveness (As) facet scores may thus account for unique variance beyond agreeableness and conscientiousness in the prediction of rating-level bias. However, given that there is a curvilinear relationship between As and leadership effectiveness (Ames & Flynn, 2007), it may be that managers who are either low or high on As may be less accurate in their ratings. We offer the following:

Hypothesis 2: Assertiveness will be negatively correlated with mean rating levels of subordinates, peers, and supervisors under varying conditions of accountability (H2a), but high and low levels of assertiveness will be related to greater inaccuracy in ratings (H2b).

Given the inconsistent results regarding conscientiousness and mean rating levels, we state no hypotheses but rather pose the following as research questions:

> *Do managers with a combination of high agreeableness, low conscientiousness, and low assertiveness make the most lenient ratings and do managers with the combination of low (or high) agreeableness, low conscientiousness, and low (or high) assertiveness make the least accurate ratings? In addition, will assertiveness levels contribute incremental validity beyond agreeableness and conscientiousness in the prediction of mean rating level and accuracy?*

The Development of a Performance Management Competency

Enlisting a motivational, goal-based theory of rating behavior that includes both cognitive and noncognitive elements suggests that a manager's attributes may affect the importance attached to the various goals that could be pursued in the context of performance appraisal activities and that the relative importance raters assign to these various goals may in turn affect the levels of

performance management competencies managers develop and sustain. We first need to define a performance management competency.

While the main focus of the scant research on rating-level bias has been on the personality characteristics of raters, especially the FFM, early research relating individual difference measures to rating outcomes focused on cognitive measures. For example, Borman (1979) found that rating accuracy and leniency were most strongly related to raters' intelligence and detail orientation. Borman and Hallam (1991) found that mastery of a particular work task was the strongest predictor of accuracy in ratings of work samples for the same task.

Hauenstein (1992) argued that rating-level bias is a complex phenomenon and, independent of a need for social approval, is a function of cognitive encoding and retrieval bias. His research suggests that there may be an underlying managerial competency made up of knowledge, skills, abilities, and motivations—a competency related to (and predictive of) rating-level bias and accuracy. The assessment of such a performance management competency (PMC) could be beneficial beyond what could be inferred from a manager's personality, past managerial performance, intelligence, or other personal characteristics. As part of a competency-modeling project, in this study a PMC was defined and a performance test was developed for its assessment. The PMC was defined as the ability to detect performance problems within the store and to take strategic action to correct them in a timely manner; the skill in the development and understanding of precisely defined performance standards; and knowledge of and imposition of these standards in the observation and assessment of performance. Detailed information on the PMC test is described in the Methods section.[3]

Another purpose of this study was to provide evidence for the criterion-related validity of the new measure of PMC. We examined whether PMC scores were related to mean rating levels and rating accuracy. In addition, we assessed the incremental validity of the PMC in the prediction of rating-level bias and accuracy beyond measures of managers' personality characteristics. We thus predict the following:

Hypothesis 3: PMC will be negatively related to rating-level bias (H3a) and positively related to rating accuracy (H3b).

Hypothesis 4: PMC will add incremental validity in the prediction of rating-level bias (H4a) and accuracy (H4b) beyond the managers' personality characteristics.

The Role of Rater Accountability in the Relationship Between Raters' Personal Characteristics and Rating Bias

Schlenker, Britt, Pennington, Murphy, and Doherty (1994) defined accountability as "being answerable to audiences for performing up to certain prescribed standards, thereby fulfilling obligations, duties, expectations, and other charges" (p. 634). Mero, Guidice, and Brownlee (2007) found that audience characteristics influence rating quality (i.e., raters accountable to a higher-status audience provided more accurate ratings, whereas those accountable to a lower-status audience provided more inflated ratings.

Mero et al. (2007) also found that raters who were required to account for and defend their ratings in a discussion with those who were rated provided more positive indicators of behavior when accountable to a lower-status audience. Our study allows for an investigation of the effects of these accountability factors on the relationships between raters' personal characteristics, including performance management competence and rating-level bias.

While the main focus of the scant research on rating-level bias has been on the personality characteristics of raters, especially the FFM, early research relating individual difference measures to rating outcomes focused on cognitive measures.

Are More Agreeable Raters More Lenient Under Conditions of High Rater Accountability?

Bernardin et al. (2000) found more agreeable raters were more lenient when they were aware that ratees knew the source of peer ratings. Yun, Donahue, Dudley, and McFarland (2005) found that more agreeable raters who anticipated face-to-face feedback sessions were more likely to inflate their ratings than less agreeable raters. They also found a significantly weaker relationship between agreeableness and rating level when a face-to-face session was not anticipated.

Our data included four sets of performance ratings that differed on the level of rater accountability, including what appears to be the two most critical "accountability" variables (i.e., the conditions in which those who are rated have specific knowledge of the rating source and an opportunity to discuss ratings with the rater). We thus were able to explore the role of appraisal accountability as a moderator in the relationships between rater characteristics and rating outcomes. Based on the results of the previous research, we hypothesize as follows:

Hypothesis 5: Rater accountability will moderate the relationship between raters' agreeableness and rating bias such that a significantly higher positive correlation will be found between raters' agreeableness and rating-level bias under conditions of high rater accountability compared to the correlation under conditions of low rater accountability.

Are Less Conscientious Raters More Lenient Raters Under Conditions of High Accountability?

Accountability has also been proposed as a moderator in the relationship between conscientiousness (C) and rating bias. Tziner, Murphy, and Cleveland (2002) found that supervisors who were less conscientious rated more leniently when they anticipated meetings with those rated, while raters who were more conscientious were not more lenient under this condition. Roch, Ayman, Newhouse, and Harris (2005) also found that raters with low levels of C tended to be lenient but only when they were aware that ratings could be directly linked to them. Raters with high levels of C did not inflate ratings even when they could be identified. However, Bernardin et al. (2000) and Bernardin et al. (2009) found that more conscientious raters were less lenient (and, in Bernardin et al., 2009, more accurate) regardless of the level of rater accountability.

The primary purpose of this study was to investigate the proposition that a manager's personal characteristics are related to rating-level bias and rating accuracy.

Assertiveness, PMC, Rating Bias, and Rater Accountability

Extreme scores on the assertiveness facet may be a more consistent predictor of rating level across different levels of rater accountability (De Dreu, Weingart, & Kwon, 2000). However, accountability may still moderate the relationship between As and rating bias such that As is a stronger predictor under conditions of high accountability where raters anticipate future interactions and potential conflict with ratees. Although we state no hypothesis, we also examined the role of accountability in the relationship between As and rating level.

In summary, the primary purpose of this study was to investigate the proposition that a manager's personal characteristics are related to rating-level bias and rating accuracy. We developed and assessed the validity of a PMC measure as a predictor of rating-level bias and accuracy. We examined whether managers' personalities and PMC levels are related to managers' mean rating levels and rating accuracy in ratings of subordinates, peers, and managers under conditions that varied on the level of rater accountability.

Methods

Participants

We collected data related to 125 associate store managers (ASMs) employed by a Fortune 500 retailer. All ASMs had worked for the retailer as ASMs for at least one year. The ASMs were: 64% male; 60% white; 13% black; 14% Hispanic or Latino; 5% Asian and 1% other.

Procedure

All ASMs participated in an assessment center (AC) designed to provide both a measure of readiness to assume a store manager job and for developmental feedback. The AC was developed based on a competency-modeling project conducted by the company and described in detail in Hagan, Konopaske, Bernardin, and Tyler (2006). All ASMs completed the "Employee Discussion" performance test, designed to assess the "Performance Management Competency."

Measurement of Personality

All ASMs completed the NEO-FFI (Costa & McCrae, 1992). The NEO-FFI is a 60-item version of the NEO Personality Inventory-Revised that provides a "brief, comprehensive measure of five domains of personality" (Costa & McCrae, 1992, p. 11). All ASMs also responded to the eight items that define the "assertiveness" facet on the NEO-PI-R (Costa & McCrae, 1992). Coefficient alphas from our study are reported in Table I.

Measurement of the Performance Management Competency

The "Employee Discussion" (ED) exercise was designed to assess (and develop) the PMC of managers. The ED was written after the PMC was defined as part a job analysis project. For the ED exercise, each manager played the role of a new store manager and was given detailed store information and asked to interview a store associate. The interview was observed by two trained assessors. Each ASM then prepared a report with recommendations for action. The two assessors then reviewed the report, interviewed the ASM, and provided independent assessments of each candidate's PMC on seven-point behavioral expectation scales (BES) derived from the job analyses pursuant to guidelines described in Bernardin and Smith (1981).[4]

Criterion-related validity of the PMC

The correlation of the two assessor ratings on PMC was .62 ($p < .01$).[5] We used the mean rating by the

TABLE I Means, Standard Deviation, Correlations and Reliabilities

		Mean	S.D.	1	2	3	4	5	6	7	8	9	10	11	12
1	Agreeable	32.18	5.63	.77											
2	Conscient	35.30	5.91	.16	.79										
3	Extravert	27.99	7.50	–.03	.28	.85									
4	EmoStab	18.13	6.89	–.21	–.41	–.17	.78								
5	Openness	25.96	6.59	–.03	.11	.21	–.09	.89							
6	Assertive	17.93	4.97	–.34	.27	.31	–.07	.10	.73						
7	PMC	4.10	1.17	–.17	.08	–.02	.03	.08	–.03						
8	HiAccTD	5.42	.78	.32	–.16	–.13	–.01	–.03	–.24	–.25					
9	LoAccTD	4.98	.67	.12	–.12	–.05	–.02	–.11	–.08	–.21	.65				
10	LoAcc-PeerMR	5.14	.78	.18	.01	–.08	–.02	.05	–.24	–.18	.21	.09			
11	UpwPAofSM	.24	.91	.15	.10	.10	–.13	–.06	–.09	–.19	.12	.00	.14		
12	PeerCA	.11	.31	–.06	–.02	–.09	.07	.03	–.11	.21	–.17	–20	.05	.03	
13	PeerElevation	.87	.49	–.01	–.00	–.14	–.03	–.24	–.03	–.27	.18	.17	.01	.23	.11

Notes: Values greater than .27, .21, or. 16, significant at .001, .01. or 05 levels of significance (1-tailed tests);
Alpha coefficients are presented on the diagonal for the FFM domains and Assertiveness.
Agreeable = Agreeableness; Conscient = Conscientiousness; Extravert = Extraversion;
EmoStab = Emotional Stability; Openness = Openness to Experience; Assertive = Assertiveness;
PMC = Performance Management Competence;
HiAccTD = Highly accountable, top-down performance appraisals;
LoAccTD = Low accountability ratings, top-down PAs from validation study;
LoAccPeerMR = Low accountability peer ratings from the MSA;
UpwPAof SM = AbsDiff upward appraisals of Store Mgrs;
PeerCA (correlational accuracy)
Peer Elevation = average of the AD scores between any ASM's particular peer rating and the "true score" rating on the same peer

assessors to test hypotheses regarding the PMC. This mean assessor rating on the PMC was correlated with assessors' consensus-derived "Overall Assessment of Managerial Potential" (r = .55, p < .01), the last mean peer rating (r = .34, p < .01), and managers' ratings of promotional "readiness" (r = .39, p < .01).

Rating Criterion Data and Conditions of Rater Accountability

We had four sets of performance ratings that were made by all of the ASMs. The rating conditions differed on their relative level of rater accountability.[6]

We classified (and validated) one set of ratings as "high rater accountability," one as "low accountability," and two other sets as "relatively low on rater accountability."

1. High Accountability, Subordinate Performance Appraisals

   For the "high-accountability" condition, we retrieved the required performance appraisals (PAs) that the ASMs had completed on their subordinates ("associates") as a part of their duties as ASMs. All ratees were aware of the source of the rating and all had an opportunity to participate in a PA interview with the ASM regarding the PA. There was no formal evaluation of the ASMs regarding how they conducted their PAs. All PAs were made on seven-point, Likert-type rating scales on (up to) five factors plus a summary rating of effectiveness. We used the summary rating to calculate the mean rating level for each rater. The rated "associates" occupied a variety of positions for the retailer, but there was substantial overlap in the position descriptions for the individuals rated across ASMs.[7]

Mean Performance Ratings and Objective Performance Data

We had access to a number of objective store and personnel measures. Among these were store sales, profit, shrinkage, employee costs/sales ratios, associate selection test scores, yield ratios, and turnover/absenteeism records. We also used a composite measure of store performance that the company used as the overall measure of store performance. None of the correlations between any of these performance measures and the mean performance ratings by individual ASMs (for the same period of time) were found to be statistically significant.

2. Low Accountability, "Top-Down" Ratings

   The second set of "top-down" ratings we classified as "low accountability" were ratings made by these same ASMs on some of their former and current subordinates as part of a criterion-related validation study for the entry-level sales position. All of the applicable ratings for the validation study were made on the same scales described earlier. We derived the average rating made by each ASM across factors. ASM ratings completed as part of the validation study were not known to those who were rated. Raters were instructed to approximate a normal distribution of ratings for the ratees whom they were evaluating. The mean number of ratees per ASM was 7.4 (SD = 2.1).

3. Low Accountability, Peer Ratings From a Multirater Appraisal System

   Another rating situation we classified (and validated) as "low accountability" called for peer appraisals completed by the ASMs (across one to four years) as part of a developmental, multirater appraisal (MRA) system used annually by the company. ASMs completed an overall rating of "promotional readiness" for peers with whom they worked using a seven-point behavioral expectation scale (BES). The average number of different peers who were rated by each ASM was 6.6 (SD = 2.1). ASMs made these ratings with the assumption that their particular ratings were confidential and that the peers they were rating could not link ratings to a particular rater. Peers received only aggregated feedback and no rater names.

   Because we had access to all data from the MRA system, we were able to derive comparison data that could be used to calculate accuracy measures. Following Murphy and Cleveland (1995), we derived a theoretical "true performance score" for each peer rated by calculating the mean rating across all raters who participated in the MRA system (except self-ratings). Thus, we derived a mean for each peer/ratee based on the mean of the subordinate, peer, and supervisory mean ratings.[8]

   Using these "true score" means, we calculated two measures of accuracy for each ASM: (1) as recommended by Murphy and Cleveland (1995), an "elevation" (E) measure was derived that was the average of the absolute difference scores between any ASM's particular peer rating and the "true score" rating on that same peer;[9] and (2) following the recommendations of Borman (1979) and others (e.g., Bernardin et al., 2009), a correlational accuracy (CA) statistic for each rater was derived by averaging the correlations of each rater's mean peer ratings (on each peer) with the grand mean rating on each of the same peers who were rated by each ASM.

4. Low Accountability, Upward Appraisals From a Multisource Appraisal System

   We also had access to the upward appraisals completed by the ASMs as part of the MRA. While most ASMs (N = 107) rated only one store manager, we used the mean of mean absolute deviations to derive the measure for the 18 ASMs who provided upward appraisals on two (or more) store managers. We then used a mean deviation score for each ASM to reflect the extent to which a particular rater's ratings deviated from the grand mean of all other ratings made of the same store manager from the MRA data across (up to) six competencies.[10]

   Because in most cases we had ratings of only one supervisor by the ASM (and not all competencies had to be rated), the one measure of accuracy available was the "elevation" measure, the average of the absolute difference score for each ratee/competency combination (across all competencies rated by the ASM for a particular boss). The absolute differences were then summed and an average derived per ASM. A relatively higher mean deviation score reflected ratings that were less accurate (but not necessarily more or less lenient) than all other appraisals made of the same manager during this same period of time. Previous research indicates that this particular deviation score reflects rater bias and is correlated with other measures of rating accuracy (Bernardin et al., 2009).[11]

Results

Table I includes the descriptive data and all intercorrelations among the major variables under study. There were no significant correlations between the demographic data pertaining to raters and any of the criterion data. The means for the conscientiousness and the extraversion FFM factors were elevated relative to published and applicable norms (Costa & McCrae, 1992). The pattern of intercorrelations among the FFM traits (and the assertiveness facet) did not differ significantly from the published matrices (Costa & McCrae, 1992, pp. 100–101).

The Stability of Rater Mean Rating Levels

We found reliable evidence for a "perceiver effect" across rating situations. For example, when we

removed all common ratees from the "top-down" data for the high- versus low-accountability conditions and then correlated ratings made on one group, all of whom were doing different jobs than the sales associates who were rated two or more years later as part of the low-accountability (validation) condition, we still found a significant positive correlation in mean rating levels ($r = .37$, $p < .001$). Even the mean subordinate ratings in the high-accountability condition were positively correlated with the mean peer ratings from the MRAs ($r = .21$, $p < .05$).

Mean Rating Levels and Rating Accuracy

We also found that managers' mean rating levels of their "subordinates" under the low-accountability condition were significantly correlated with two measures of rating accuracy derived from other ratings made by these same raters. The mean subordinate ratings in the high-accountability condition were positively correlated with the peer "elevation" ($r = .18$, $p < .05$) and CA measures ($r = –.17$, $p < .05$).

Rater Personality, PMC, and Rating Outcomes

We first report the analyses related to hypotheses followed by the research questions and ancillary findings. We found some support for the predicted positive relationships between ASM agreeableness (A), rating level, and rating accuracy. In support of H1a, ASM A scores were positively correlated with subordinate mean rating levels under the high-accountability condition ($r = .32$, $p < .01$) and also mean rating levels of ASM peers ($r = .18$, $p < .05$) and supervisors ($r = .16$, $p < .05$). However, the correlation between (A) and the mean subordinate rating level for the low-accountability (validation) condition was not significant ($r = .12$, $p = .06$). In support of Hypothesis 1b regarding a U-shaped relationship between A and E, ASMs who were either high or low on A (over 1.5 SD above/below the mean) were the least accurate compared to the grand means of ratings on peers ($p < .01$) and supervisors ($p < .05$).

In support of Hypothesis 2a, assertiveness (As) was negatively correlated with the mean rating level of subordinates under the high-accountability condition ($r = –.24$, $p < .01$) and also when ASMs were rating their peers as part of the low-accountability MRA process ($r = –.24$, $p < .01$). In support of Hypothesis 2b, we found that ASMs with either high or low levels of As were more inaccurate compared to ASMs who were less extreme on As ($p < .05$). Using 1.5 SD as the criterion, we found that ASMs who were either high or low on As were also the least accurate in their ratings compared to the grand means for both the peer and supervisory ratings ($p < .05$).

Hypothesis 3 proposed that ASM scores on the PMC would be correlated with rating level (H3a) and (H3b) rating accuracy. In support of H3a, PMC assessments were significantly correlated with ASMs' subordinate ratings made under conditions of high accountability ($r = –.25$, $p < .01$), with the subordinate ratings under conditions of low accountability ($r = –.21$, $p < .05$), and also with the peer ratings under low-accountability conditions ($r = –.18$, $p < .05$). In support of H3b, the PMC measure was significantly correlated with the E measure of accuracy for the ASMs' upward appraisals ($r = –.19$, $p < .05$) and both accuracy measures for the peer appraisals ($r = –.27$, $p < .01$ for E and $r = .21$, $p < .01$ for the CA measure).

Hypotheses 4a and 4b predicted that scores on PMC would add incremental validity to the prediction of rating level and accuracy beyond the contributions of the personality measures. To test these hypotheses, we performed two hierarchical regression analyses using the high-accountability, subordinate ratings as the dependent variable to test H4a and the peer E measure to test H4b (see Table II). We entered rater scores on A, C, and As in the first step and the mean PMC measure in the second step. In support of both H4a and H4b, results indicated that PMC scores contributed incremental validity beyond raters' personality variables in predicting both mean subordinate ratings ($p < .05$) and peer rating accuracy ($p < .05$).

Rater Accountability as a Moderator

Hypothesis 5 stated that rater accountability would moderate the relationship between raters' A and rating level such that a significantly higher correlation would be found between raters' A and rating level under conditions of high rater accountability compared to the correlation between A and mean rating level under conditions of low accountability. We restricted comparisons between the high- and low-accountability conditions to the subordinate ratings. The best test of H5 was to compare ASMs' ratings of the same subordinates under the high- versus low-accountability conditions. Under the high-accountability condition, A was significantly correlated with the mean rating level ($r = .32$, $p < .001$). However, the correlation between A and ratings on the same individuals for the low-accountability condition was not significant ($r = .12$, $p = .06$). In support of H5, the difference in these two correlations was statistically significant ($p < .01$).

TABLE II Results of Hierarchical Regression Testing the Independent Variables with Controls

	Low ACC Peer Ratings		High ACC Subordinate PA		Peer Rating Elevation	
Hypothesis	**H2b**		**H4a**		**H4a**	
Step and Variable	**1**	**2**	**1**	**2**	**1**	**2**
Conscientious	–.02	.05	–.19*	–.20*	.02	.02
Agreeableness	.18*	.10	.33***	.28**	–.03	–.09
Assertiveness		–.22*	–.08	–.10	–.05	–.08
PMC				–.22**		–.29**
R^2	.03	.07	.16	.20	.00	.08
R^2 *change*		.04		.04		.08
F	2.00	2.95	7.43	7.50	.08	2.70

*p <.05
**p < .01
***p < .001
[a]Reported values are standardized regression coefficients
N = 125
Low ACC Peer Ratings = mean peer ratings in condition of low accountability
High ACC Direct Report Ratings = mean ratings of subordinates under condition of high accountability
Low ACC Peer Elevation Ratings = mean peer elevation ratings in condition of low accountability (accuracy of ratings)
PMC = performance management competency

Research Questions and Ancillary Analyses

Raters' C scores were negatively correlated with their mean rating levels of subordinates but only under the high-accountability condition (r = –.16, p < .05), indicating that more conscientious managers rated less leniently. No other correlations between C and criterion measures were significant (see Table I).

We found evidence that raters with high levels of A, low levels of C, and low levels of As would be the most lenient of raters in the high-accountability condition. While we identified only 10 ASMs who matched this particular personality profile, the mean subordinate ratings from this group was 6.17 (SD = .47); significantly higher (p < .01) than the mean rating of all other raters (5.35; SD = .77).

We also examined whether the combination of high A and low C would have a higher correlation with rating level in the high- (versus low-) accountability condition. Comparing regressions using the mean ratings of the same subordinates, the adjusted R^2 in the prediction of highly accountable ratings of "subordinates" using A and C as predictors was .14 versus .02 in the prediction of the subordinate ratings under the low-accountability condition.

We also identified ASMs who had the following scores on the FFM measures: either high or low scores on A or high or low scores on As and low scores on C. Compared to others, elevation accuracy measures for these 16 ASMs were significantly more inaccurate when rating both peers (p < .01) and supervisors (p < .05).

Scores on the As facet of the FFM contributed incremental validity beyond A and C in the prediction of rating level for the peer ratings. Using hierarchical regression, we found that the As facet contributed significantly to the prediction of mean peer rating level (see Table II). However, As did not contribute significantly (beyond A and C) in the prediction of the mean ratings from the high-accountability, subordinate ratings.

Testing Overall Fit

To determine the overall fit of our hypothesized relationships, we constructed two path analytic models using LISREL. Figure 1 presents the first model with A, C, As, and PMC as the predictors of matched subordinates in conditions of both high and low accountability and as the predictors of peer ratings under low accountability. Figure 2 presents the second model with the same personality characteristics and PMC as the predictors of rating elevation and CA. Table III presents the results and fit indices.[12] Fit indices for both models indicated acceptable fit (Hu & Bentler, 1998; Kline, 1998; Krishnan & Singh, 2010).

Significant paths between the PMC and all outcome variables highlight the validity of PMC in predicting rating-level bias under varying conditions of accountability and rating accuracy definitions. Significant paths between A and C with the ratings of subordinates were only factors found under the highly accountable condition while only A was significant under the low-accountability condition for peer ratings.

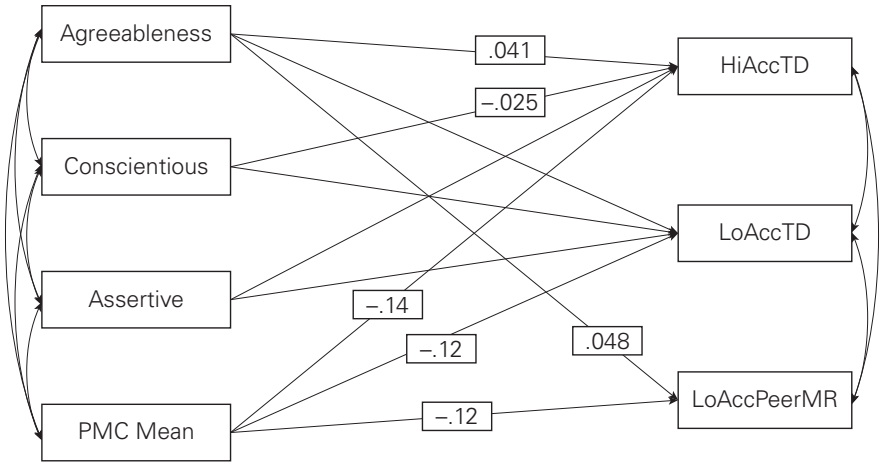

FIGURE 1. Path Analytic Model with Rating Level of Subordinates and Peers at Varying Levels of Accountability as Outcome Variables

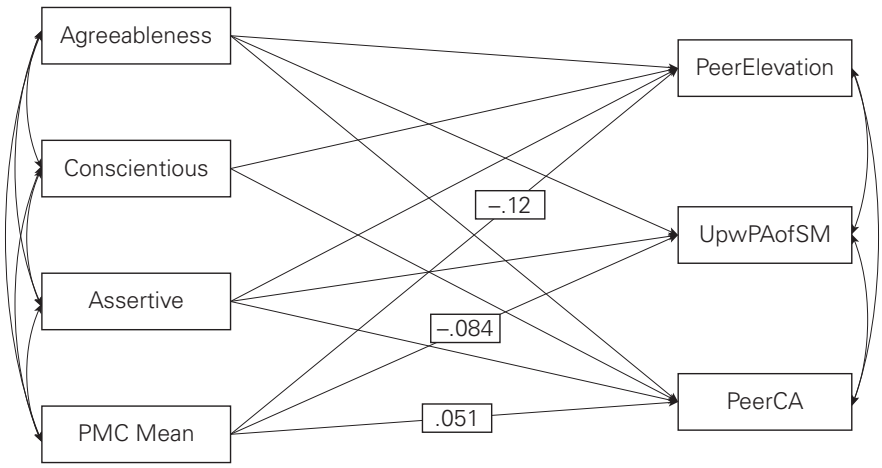

FIGURE 2. Path Analytic Model with Accuracy Measures of Peers and Supervisors as Outcome Variables; Standardized Beta Weights of Significant Paths

TABLE III **Results of Two Path Analytic LISREL Models**

Model	Fit Measures
1	Normed $X^2 = 2.77$
	SRMR =.029
	CFI = .97
	GFI = .99
	NFI = .98
	RMSEA = .057
2	Normed $X^2 = 1.17$
	SRMR = .017
	CFI = 1.00
	GFI = 1.00
	NFI = .98
	RMSEA = .038

Curvilinear relationships in other hypothesized paths probably mitigated significant effects.

Discussion

Our results extend the provocative findings from Wood et al. (2010) regarding both "perceiver effects" and "positivity effects" in the ratings of others and the findings of Kane et al. (1995) on the stability of mean rater rating levels across very different rating situations. We also partially replicated the findings from Bernardin et al. (2009), who found that "individuals who are both more agreeable and less conscientious represent the worst combination of rater traits for effective ratings" (p. 309). When we added the assertiveness facet to the predictive equation, we found that we could improve the prediction of mean rating level with this additional personality variable from the FFM. In addition, we found support for

the validity of a curvilinear relationship between levels of both agreeableness and assertiveness and rating accuracy. While our results clearly support Guilford's (1954) "dispositional theory" of rating, we also found rater accountability was a moderator in the relationship between A and mean rating levels. More agreeable supervisors are more likely to rate leniently under the most typical appraisal situations in which subordinates are aware of the rating source and may discuss the appraisal with the rater.

We also found strong support for the validity of a performance management competency measure, scores on which were correlated with both mean subordinate and peer rating levels and with measures of rating accuracy for both peers and supervisors. PMC scores contributed incremental validity beyond raters' personalities in predicting mean rating levels and accuracy.

Managers' mean ratings of past performance of their subordinates were positively correlated with their mean PAs of their bosses and the promotional readiness of their peers. Our findings provide further support for the "perceiver effects" shown by Kane et al. (1995) and Wood et al. (2010). We found the effect for managers rating their colleagues' job performance or potential across three hierarchical levels of an organization. Managers who tended to rate the performances of their subordinates more positively (or negatively) also tended to rate the performances of their bosses and peers more positively (or negatively) as well. Despite the differences in the various rating situations studied here and even as much as a three-year period of time between when many of the ratings were made, we still found that managers' mean rating levels across these appraisal situations were correlated. We have identified a rater bias effect indicating that raters who are more lenient (or harsh) tend to rate this way over time and circumstances with an average correlation of .21 among means across all mean comparisons (excluding the set of the subordinate ratings involving the same ratees). This .21 correlation can be considered an estimate of rater bias effects that could have deleterious effects on the validity and undermine the usefulness of performance appraisal data if ratings are compared across raters and no adjustments are made based on the tendencies of particular raters. Research is needed to determine if such adjustments would enhance the validity of ratings.

So how do we account for a rater's "dispositions"? We found managers' levels of A, C, As, and PMC were correlated with the mean ratings of their subordinates for the most typical appraisal circumstances. We thus partially replicated and extended previous research involving students rating other students (Bernardin et al., 2000; Bernardin et al., 2009; Wood et al., 2010). We found that self-report measures of personality traits, especially A and As, were predictive of managers' mean rating levels. Not only did managers' A scores correlate (positively and significantly) with their mean rating levels across all but one rating circumstance, but A was also shown to be related to rating accuracy when managers rated their peers and their bosses. Guilford (1954) was apparently correct about a rater's stable "disposition" having an effect on rating tendencies across circumstances.

The correlation between A scores and the mean subordinate ratings under conditions of low accountability was not (quite) statistically significant (r = .12). We believe this finding may be related to the different goals more agreeable managers may have had across these two very different rating situations that required subordinate ratings. For the "high-accountability" subordinate rating condition, more agreeable managers were probably more influenced than other managers by "harmony" goals in dealing with their subordinates, avoiding conflicts or discomfort with negative ratings (Villanova, Bernardin, Dahmus, & Sims, 1993), and also pleasing the recipients of the appraisal (Randall & Sharples, 2012). These goals were not salient in the low-accountability subordinate rating condition. Here, raters were aware that the completed ratings would not be revealed to the subordinates who were rated. In addition, relative to other managers, perhaps more agreeable managers may have been relatively more compliant with the administrative instructions provided to raters as part of the criterion-related validation study. These instructions included direction for raters to "approximate" a normal distribution of ratings. Mean ratings by the most agreeable raters were significantly higher under high-accountability conditions compared to ratings of the same subordinates under the low-accountability condition. These same (more agreeable) individuals were also more likely to rate peers and supervisors more leniently and less accurately.

One provocative finding was the curvilinear (U-shaped) relationship we found between A and the elevation measure of accuracy. While the relationship between A and mean rating level was linear, managers who were either very low or very high on A made peer and upward ratings that deviated the most from the grand mean ratings on the same individuals. Our results suggest two inflection points for scores on A if a primary administrative interest is in facilitating more accurate appraisals. Managers closer to the mean on A were the most accurate raters using the elevation

measure, what is generally considered to be the most important measure of rater accuracy for personnel decisions (Murphy & Cleveland, 1995, p. 288).

In this first study that involved a facet underlying the FFM model of personality as related to rating behavior, we also found strong effects for the value of the As facet as a predictor of rating level and rating accuracy. Raters who are less assertive raters rated more leniently when rating subordinates in both the high- and low-accountability conditions and also when rating their peers and supervisors. Also, like the findings for A, ASMs who were either at the high or low end of As tended to be more inaccurate compared to the ratings from ASMs who were closer to the average level on this facet. These results provide support for differentiating the underlying facets of extraversion and the divergence of these facets of extraversion (DeYoung et al., 2007). Assertiveness was also part of the personality profile of the most lenient raters under conditions of high accountability. Given the sample size for the profile analysis, our results regarding combinations of traits and competency are merely suggestive. More research is needed on combinations of traits and facets and the determination of inflection points for low and high scores on A and As.

Our most significant finding was that scores on the new measure of performance management competence were correlated with rating-level bias and rating accuracy. The PMC measure was derived from a competency-modeling project for the retail store manager's job. For the development of the PMC performance test, we relied on past research investigating predictors of accuracy in rating (e.g., Borman, 1979; Funder, 2012; Hauenstein, 1992) and the results of the job analysis. Store managers and district managers who participated in the development and validation of the PMC measure regarded the defined PM competency as a critical and essential component for a store manager's success. The significant correlations of scores on the PMC with mean rating levels and accuracy provide strong evidence for the criterion-related validity of the PMC. In addition, the PMC measure contributed incremental validity in the prediction of mean rating level and accuracy beyond the contributions of A and As. Our results suggest that managers who score low on the PMC who are also nonassertive, highly agreeable, and with low levels of C are the most likely to rate leniently. Our research also suggests they may rate less accurately.

Contrary to the recent findings of Randall and Sharples (2012), we found that rater accountability moderated the relationship between agreeableness and rating level. Agreeableness was more strongly related to mean rating level under conditions of high (versus low) rater accountability. As predicted, more agreeable managers had a tendency to rate more leniently in the most typical appraisal situations in which those who are rated received feedback on administratively significantly ratings, were aware of the rating source (their boss), and have an opportunity to discuss the ratings with their boss.

It is likely that more agreeable raters have performance management and appraisal goals that tend to place more emphasis on seeking to establish and maintain convivial working relationships with coworkers, avoiding conflicts and confrontations, and pleasing the recipients of their feedback. More agreeable managers may be more inclined to rate leniently under conditions of high accountability, where these goals are more salient (Randall & Sharples, 2012). More agreeable managers may place more importance on "harmony" versus "fairness" goals and, consequently, rate more leniently and less accurately (Wong & Kwong, 2007).

Our results suggest that managers who score low on the PMC who are also nonassertive, highly agreeable, and with low levels of C are the most likely to rate leniently. Our research also suggests they may rate less accurately.

Hauenstein (1992) argued that if administrators or managers convey no suggestions or stipulations that relatively higher ratings must be justified or that particular aggregated rating distributions should be obtained, then the most "cognitively efficient strategy" for raters is to render evaluations that fit the likely expectations of the constituent they are most concerned about satisfying. Thus, in the common scenario in which an organization places little or no emphasis on rating levels or distributions, raters are more likely to be influenced by ratees' expectations of high ratings. Raters who are high on agreeableness and low on assertiveness may be relatively more concerned with these expectations. But as evidenced by the lower correlation between A and mean subordinate ratings in the low-accountability condition, more agreeable raters may pay close attention to clear expectations conveyed by higher authorities.

This is the first study involving managers rating their subordinates that investigated both rating level and rating accuracy. Raters' personality traits, especially A, and PMC were not only related to rater rating level; agreeableness and PMC scores were related to measures of rating accuracy. We also found that both high and low levels on A were related to accuracy when using an elevation

measure that best represents the overall accuracy of each rating. Using the grand means from the MRA data as "true score" estimates, we found that the ratings from both highly agreeable and highly disagreeable managers tended to deviate the most from these "true scores." In addition, the levels of deviation were not significantly different (i.e., highly disagreeable ASMs were no more inaccurate than ASMs who were highly agreeable).

When we used mean ratings from the highly accountable, subordinate ratings, our results essentially replicated and extended the most recent research findings showing that a profile made up of the most agreeable and least conscientious raters comprised the worst combination of rater FFM traits for effective ratings (Bernardin et al., 2009). In addition, including a low As and a low PMC score with this profile increased our ability to predict mean rating level.

When we used mean ratings from the highly accountable, subordinate ratings, our results essentially replicated and extended the most recent research findings showing that a profile made up of the most agreeable and least conscientious raters comprised the worst combination of rater FFM traits for effective ratings.

In the highly accountable rating situation, managers were aware that their subordinates knew the source of the ratings. Each ratee received a feedback report that included the manager's ratings and signature. In addition, both raters and ratees were aware that the ratings had administrative significance for the associates. In this very common rating situation, managers with low levels of conscientiousness and PMC plus high levels of agreeableness had the highest mean rating levels. These higher-than-average rating levels could not be explained by any objective data available that could have supported an argument that mean rating-level differences were correlated with legitimate differences in true unit-level performance across managers. PMC scores also provided incremental validity over the personality variables in the prediction of mean subordinate rating level in the highly accountable condition.

We also found new and unique evidence for nonlinear relationships among levels of A and As and measures of rating accuracy. Both high and low levels of A and high and low levels of As were related to less accurate ratings. Compared to ratings by other raters in ratings of peers and supervisors, performance ratings by managers with extreme scores on A or As were the most distant from the grand mean rating levels for both peers and upward appraisals. High and low levels of A or As are related to greater rating inaccuracy.

Our two path-analytic models provide support for the predictive relationships among the personality characteristics, the PMC, and rating-level bias under varying conditions of accountability and across hierarchical levels. In addition, our model regarding these same predictive variables was sustained when we used measures of rating accuracy with the peer and upward appraisal data.

Our results corroborate theory in the context of past research on the motivations and proclivities of very agreeable managers (Bernardin et al., 2009; Wood et al., 2010). The findings also suggest a cautionary note about the ability of very agreeable managers to supervise even the least effective of workers. Yun et al. (2005) found that highly agreeable raters were more likely to rate leniently in a condition of high rater accountability to the ratee where they anticipated a face-to-face feedback session with a ratee of high or moderate performance but their ratings for a ratee with low performance were significantly lower than those provided by raters who were low on A. We found no support for this finding. More agreeable raters also rated the least effective peers more leniently. In addition, managers who were relatively more agreeable were also the least likely to recommend probationary action for the substandard performance portrayed in the script for the PMC performance test (the action considered most effective by the assessors).

Strengths, Limitations, and Future Research

Our study had several strengths. We studied real managers evaluating their subordinates, peers, and supervisors. This is a rare circumstance for research on rating bias and rating accuracy in particular. Almost all published research investigating the role of individual difference variables in the prediction of rating bias and accuracy has involved students rating other students or contrived rating situations involving the rating of videotaped or scripted performances. Our research also allowed for the study of an organization's actual performance appraisals completed over several years, across different circumstances and different hierarchical rating levels.

We believe this is also the first field study that used MRA data to define "true scores" for the calculation of accuracy measures. As recommended by Murphy and Cleveland (1995, pp. 290, 291), we derived mean ratings on peers and managers across rating sources (except self-ratings) and used these means as representations of "true scores." With these data, we found significant relationships between accuracy estimates for individual raters and their dispositional attributes.

Although we studied only one organization, an advantage is that other possible systematic factors that could have an impact on ratings were controlled. We know that the social context of appraisal can have an impact on ratings and rating behavior (Levy & Williams, 2004). These contextual variables are largely held constant in our study. In addition, our ability to examine the impact of accountability factors was enhanced by controlling relevant factors such as the particular ratees and the conditions for data collection. More research is needed to test whether other factors may be related to rating level and accuracy. Future research should attempt to replicate our findings in other organizational settings and with other individual difference measures such as the level of rater discomfort (Villanova et al., 1993).

Another important limitation of our work is that our methods precluded establishing a direction of causality for relationships between a manager's personality and his or her ratings of coworkers. For example, more agreeable people are generally friendlier than others (Costa & McCrae, 1992). Does this friendliness foster a more favorable working relationship with all associates (subordinates, peers, and supervisors) and thus help explain the higher ratings bestowed by these more friendly managers across organizational levels? For the personality factors, we believe the causal effect probably goes in both directions. There can be little doubt that the tendency to be highly agreeable or highly assertive can have an impact on the behavior of others (for better or for worse). While we correlated unit-level, objective outcomes measures with the mean subordinate ratings and found no reliable justification that could justify differences in managers' mean ratings in terms of higher (or lower) outcomes, we do not regard the available objective data as ideal for validation or understanding the direction of causality.

The levels of rater accountability represented in the four data sets studied here are particularly compatible with accountability factors relevant to appraisal systems used in organizations, including the peer and upward appraisals that are part of multirater appraisal systems. Despite the strong effects for agreeableness and assertiveness, more research is needed on rater attributes in the context of different rating conditions and also with a focus on the differences in goals that raters/managers have for performance appraisal and management (Wang, Wong, & Kwong, 2010). For example, do more agreeable managers have goals regarding appraisal and particular ratings in particular that focus on fostering a more friendly working relationship with subordinates? How would more agreeable managers react to conditions of high accountability to administrators reviewing rating distributions and tendencies? Our findings regarding how the more agreeable raters acquiesce to rating instructions in the low-accountability condition indicate priming or training may counteract ratees' expectations of higher ratings.

Future research should examine other FFM facets underlying the five FFM domains and, in particular, the six facets underlying A. The 60-item version of the NEO Five-Factor Inventory was used in this study, but this instrument sacrifices information regarding the narrow facets that define each FFM domain. The assertiveness facet was used primarily because research indicated its potency in predicting behavior in confrontational situations. Future research should examine the relative predictability of other facets or item cluster subcomponents. Chapman (2007) found that subcomponents of the Five-Factor Inventory could enhance prediction when validation involved more narrowly focused but important criteria. Managerial appraisal functions and dealing with difficult performance problems could be examples of more carefully defined criteria that could be more successfully predicted with facets or subcomponents of the FFM.

Our findings regarding how the more agreeable raters acquiesce to rating instructions in the low-accountability condition indicate priming or training may counteract ratees' expectations of higher ratings.

Research on narrower trait subcomponents could also potentially lead to greater understanding of the multifaceted nature of managerial performance. Barrick, Mount, and Judge (2001) reported a meta-analytic correlation of .10 for agreeableness and overall managerial performance.[13] Our study found that a higher score on agreeableness was related to more lenient (and less accurate) ratings, conditions this organization (and most organizations) would regard as problematic. Future research needs to sort out this apparent discrepancy and use other operational definitions of rating effectiveness. Le et al. (2011) did not investigate the possibility that the relationship between agreeableness and performance may be complex, and perhaps curvilinear. Our findings indicate a nonlinear relationship that might help explain this low correlation. Certainly understanding the relationships among the first-order factors and dimensional criteria that comprise broader performance constructs can benefit from efforts that focus on the predictor side with greater fidelity.

Practical Implications

Our findings have several practical implications. First, this first test of the validity of the PMC construct and performance test supports use of such a measure as a diagnostic tool for managerial training purposes and perhaps as a selection device. More research is necessary, including evidence for the construct and predictive validity of the PMC using performance measures as criteria, especially rating level, accuracy, and reactive measures (Pichler, 2012). Used as a selection device, the PMC could serve as one of a number of criteria used to determine suitability or readiness. PMC scores may indicate those who require training in order to competently conduct performance management responsibilities.

Our results also indicate that, under the most common appraisal circumstances, more agreeable, less conscientious, less assertive, and less competent managers are more likely to rate leniently and perhaps less accurately. When administratively significant ratings are conducted but there is little or no accountability for ratings to upper management, this profile appears to identify more biased raters—those in need of remedial training. We recommend assessment of these attributes along with rating data for diagnostic and training purposes. Understanding how these variables influence ratings can be included in management training programs. This would inform trainees concerning the characteristics that have been demonstrated to elicit more accurate performance ratings. Training can emphasize the importance of a performance management competency for managerial effectiveness. There are clearly certain rater and situational characteristics related to a likelihood of evaluating performance less effectively. Our goal should be to minimize the influence of these factors through training and selection.

Used as a selection device, the PMC could serve as one of a number of criteria used to determine suitability or readiness.

Our findings also underscore the need for management to hold raters more accountable for their rating distributions. Accountability may be simple to implement. Rater mean rating levels that deviate from averages for comparable rating situations should require precise performance-related justifications. The highest levels of leniency were found in our study when raters' rating data was not formally evaluated by management. Rating data for managers should be monitored. More variance and less leniency were found when raters were informed that their ratings would be reviewed and that a normal distribution of ratings was expected. If ratings that deviate from this expected distribution cannot be justified, consideration should be given to correcting/adjusting ratings made by particular raters so as to enhance the validity of the performance management system and personnel decisions that require comparisons across raters.

The incentives to provide lenient ratings may outweigh the perceived costs, particularly for those raters with low levels of agreeableness, conscientiousness, PMC, and assertiveness. Higher ratings obviously likely please recipients, helping to maintain easy working relationships. Yet when there is a propensity to inflate ratings over time and across all ratees, high-performing ratees may become demotivated as they become aware that their ratings and associated pay raises vary little from their mediocre counterparts. High performers prefer individualized rewards and recognition (Barber & Simmering, 2002; LeBlanc & Mulvey, 1998). High performers are particularly sensitive to a lack of differentiation in ratings and more likely to seek employment elsewhere if they do not feel they have been recognized with the financial rewards they feel they deserve (Rynes, Gerhart, & Minette, 2004). Furthermore, satisfaction with pay is related to organizational commitment and trust in management (Williams, McDaniel, & Nguyen, 2006). Identifying, selecting, and retaining high performers can be costly, so developing effective appraisal systems that can measure a workers' true performance is important. To that end, identifying and training managers competent at performance management is critical.

Notes

1. We suspect the contrary findings from these two studies can be explained by the rating circumstances from the Bono et al. study. Bono et al. (2012) had subordinates complete a self-report measure of conscientiousness and then evaluate their leaders. It may be that the subordinates' conscientiousness scores were affected by the performance evaluations they had received from the leaders whom they were evaluating. Of course, the ratings of leadership effectiveness would also be affected by those same performance evaluations. This form of "reciprocal leniency" is well documented (Bernardin, 1981; Clayson, Frost, & Shefflet, 2006) but was not operative in the Bernardin et al. (2009) study, where students made anonymous ratings of their peers.

2. The other reasons for including the assertiveness facet are: (1) Previous research involving the "Big Five" traits and rating behavior has used the popular 60-item version of the "Big 5" (namely, NEO-FFI). While Saucier (1998) found that almost all of the FFM facets could be measured using the NEO-FFI, the one

facet not measured reliably on this questionnaire is assertiveness. Only one item from the 240-item NEO-PI-R that is used to define the assertiveness facet is included on the NEO-FFI. (2) Our analysis of the items on the NEO-PI-R concluded that the particular items used to derive the assertiveness facet (e.g., "I am dominant, forceful, and assertive") were the most likely to be correlated (negatively) with leniency in ratings (and perhaps in a curvilinear manner with rating accuracy). (3) All other facets identified and defined using the NEO-PI-R but not represented on the 60-item version do not indicate any empirical or theoretical relationships to rating outcomes, including rating-level bias, leniency, accuracy, or (as described in the Methods section) performance management competence. (4) Assertiveness is not highly correlated with the aggregated extraversion score of the FFI and has a negative correlation with agreeableness (–.15). (5) Scores on the extraversion FFM domain are unrelated to agreeableness (r = .04; Costa & McCrae, 1992).

3. Several factors were identified by subject-matter experts as important correlates of the PMC, including a goal orientation for observing and evaluating performance that emphasized making and providing accurate assessments. In addition, one step of the job analysis (described in the Methods section) found that subject-matter experts regarded the ideal profile predictive of high levels of PMC was a manager with high levels of conscientiousness, moderate levels of agreeableness, and high levels of assertiveness. Based in particular on information derived from the critical incident method, the performance test was developed to assess a store manager's "performance management competency." This performance test was included as an exercise for the assessment center.

4. As part of the assessment center development, a content-evaluation panel (n = 11), made up of former and current store managers, judged the overlap between the PMC and the performance test based on the use of "substantive validity indices" (Anderson & Gerbing, 1991). All panel members assigned the "Employee Discussion" exercise to the predicted PMC. In addition, the same content-evaluation panel indicated that a high rating on the particular BES purported to measure the PMC reflected a high level of performance on this competency. The same panel of subject-matter experts was also asked to indicate whether they believed certain personal attributes would be predictive of high levels of the PMC. Following procedures described by Raymark, Schmit, and Guion (1997), they were given definitions (and sample items) from the FFM and certain FFM facets and asked to indicate to what extent low, moderate, or high levels of the FFM traits/facets would be predictive of low, moderate, or high levels of the PMC. As expected, these experts indicated that managers who were low to moderate on agreeableness and high on conscientiousness and assertiveness would have high levels of PMC.

5. This correlation was compiled across 24 different assessors who rated this performance test with two assessors per ratee. Assessors were randomly assigned as either rater 1 or rater 2.

6. We conducted a study to determine if the rating situations were properly classified in terms of their relative accountability. We asked eight MBA students (all of whom had experience as supervisors or managers and who had done performance appraisals) to consider each rating situation as described and to then rate each situation in terms of a rater's level of accountability. We also asked these same students to rate the extent to which the average rater would be likely to rate leniently (i.e., higher than deserved) in each rating situation. The responses to two questions supported the classification levels we describe for the accountability conditions in our study. The four accountability conditions shown to be most strongly related to rating leniency are rating situations in which: (a) raters are aware that ratees know the source of the ratings (Mero et al., 2007); (b) raters are aware that ratees will have an opportunity to discuss the ratings (Klimoski & Inks, 1990; Mero et al., 2007; Tziner et al., 2002); (c) ratings have administrative significance; and (d) raters anticipate future interactions with the ratees (Murphy & Cleveland, 1995). We had one rating situation in which each of these conditions were met (i.e., subordinate ratings of "direct-reports"), classified as "high accountability." A second set of subordinate ratings was collected as part of a criterion-related, validation study in which none of the four accountability conditions described above applied. This rating condition was classified (and validated) as "low accountability."

   These ratings were completed by the ASMs on their peers and their supervisors as part of a multisource appraisal program. The ASM/raters were aware that ratees would receive only aggregated data and would be unaware of the source of particular ratings. No discussion with ratees regarding their ratings was anticipated. Furthermore, the raters were aware that these ratings also had no administrative significance for the ratees. Finally, there was no accountability for raters in terms of an assessment of rating outcomes by persons who might be charged with such a task.

7. The rating format for overall effectiveness was the same for all associates who were rated. We derived these ratings from a database of performance appraisals for the previous two- (or more) year period prior to the ASMs' participation in the assessment center. The average number of unique ratees per rater was 14.2

(SD = 3.19). Although we conducted other analyses with subsets of these data, for most hypothesis testing, we used the mean rating across all ratings of any particular ratee to calculate the mean rating level for each rater.

8. Based on a predictive validity study, the unit-weighted mean using the mean subordinate, peer, and supervisory assessments of "promotional readiness" had the highest level of validity compared to other methods of aggregating the multisource data (Tyler & Bernardin, 2003).

9. We excluded self-ratings and subordinate ratings in the derivation of this measure.

10. We excluded the self-ratings and each particular ASM's ratings in the calculation of this grand mean.

11. We conducted a study to determine if the rating situations described were properly classified in terms of their relative accountability. We asked eight MBA students (all of whom had experience as supervisors or managers and who had done performance appraisals) to consider each rating situation as described and to then rate each situation in terms of a rater's level of accountability. We also asked these same students to rate the extent to which the average rater would be likely to rate leniently (i.e., higher than deserved) in each rating situation. The responses to two questions supported the classification levels we describe for the accountability conditions in our study.

12. We report X^2 statistics, root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), goodness of fit index (GFI), the comparative fit index (CFI), and the normed fit index (NFI). In the case of GFI, CFI, and NFI, a value of 0 indicates the worse fit, while a value of 1 indicates a perfect fit, while an RMSEA of .08 or lower indicates good fit. An SRMR of 0 indicates perfect fit.

13. Most studies included in this meta-analysis do not report tests of linearity or curvilinearity in this relationship. The low average correlation between A and managerial performance could to some extent be caused by nonlinearity at the high level.

H. JOHN BERNARDIN is Emeritus Professor of Management in the College of Business at Florida Atlantic University. He is past chair of the Division of Personnel/Human Resources of the Academy of Management and served as director of the applied psychology doctoral program at Virginia Tech. He served as coeditor of *Human Resource Management Review* and has served on the editorial boards of numerous journals, including the *Academy of Management Review*, the *Journal of Organizational Behavior*, and *Human Resource Management*. Dr. Bernardin is the author of six books and more than 100 articles, and his main research topics are staffing, performance management, and legal issues related to HRM.

STEPHANIE THOMASON is an associate professor of management and the associate dean of the John H. Sykes College of Business at the University of Tampa. She also serves on the board of the National Society of Experiential Education and chaired its 2013 Annual Conference. Her research interests include human resource selection and assessment, employee entitlement, and personality. She has had more than 25 articles published or presented in peer-reviewed journals and academic conferences. Her articles have appeared in journals such as the *International Journal of Selection and Assessment,* the *Journal of Cross-Cultural Psychology,* and the *International Journal of Innovation Management.*

M. RONALD BUCKLEY is the JCPenney Company Chair of Business Leadership and a professor of management and a professor of psychology in the Michael F. Price College of Business at the University of Oklahoma. He is interested in a variety of HRM topics, including interview decision making, fairness in selection, and organizational socialization. Dr. Buckley has written more than 120 peer-reviewed articles, many of which have appeared in the *Journal of Management, Academy of Management Review,* the *Journal of Applied Psychology, Personnel Psychology, Educational and Psychological Measurement, Organizational Behavior and Human Decision Processes,* and the *Journal of Organizational Behavior.*

JEFFREY S. KANE is president of Professional Statistical Services, located in Myrtle Beach, South Carolina. He currently provides statistical analysis, management science, and litigation support services to clients around the world. Dr. Kane worked for 23 years as a professor of industrial/organizational psychology and management at several universities.

He is the founder and founding editor of *Human Resource Management Review* and has written numerous articles in leading journals, including *Personnel Psychology*, *Psychological Bulletin*, *Academy of Management Journal*, the *Journal of Management*, and *Human Relations.* His major interests are statistical analysis, psychometrics, performance management, motivation, selection, compensation, decision making, and program evaluation.

References

Ames, D. R., & Flynn, F. J. (2007). What breaks a leader: The curvilinear relation between assertiveness and leadership. Journal of Personality and Social Psychology, 92, 307–324.

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. Journal of Applied Psychology, 76, 732–740.

Atwater, L. E., Wang, M., Smither, J. W., & Fleenor, J. W. (2009). Are cultural characteristics associated with a relationship between self and others' ratings of leadership? Journal of Applied Psychology, 94, 876–886.

Barber, A. E., & Simmering, M. J. (2002). Understanding pay plan acceptance: The role of distributive justice theory. Human Resource Management Review, 12, 25–42.

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

Bartells, L. K., & Doverspike, D. (1997). Assessing the assessor: The relationship of assessor personality to leniency in assessment center ratings. Journal of Social Behavior & Personality, 12, 179–190.

Bernardin, H. J. (1981). The effects of reciprocal leniency on the relationship between 'consideration' scores on the LBDQ and performance ratings. Proceedings of the Academy of Management, pp. 131–135.

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. Journal of Applied Psychology, 85, 232–234.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales. Journal of Applied Psychology, 66, 458–463.

Bernardin, H. J., Tyler, C. L., & Villanova, P. (2009). Rating level and accuracy as a function of rater personality. International Journal of Selection and Assessment, 17, 300–310.

Bono, J., Hooper, A. C., & Yoon, D. J. (2012). Impact of rater personality on transformational and transactional leadership ratings. The Leadership Quarterly, 23, 132–145.

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

Borman, W. C., & Hallam, G. L. (1991). Observation accuracy for assessors of work-sample performance: Consistency across task and individual-differences correlates. Journal of Applied Psychology, 76, 11–18.

Chapman, B. P. (2007). Bandwidth and fidelity on the NEO-Five Factor Inventory: Replicability and reliability of Saucier's (1998) item cluster subcomponents. Journal of Personality Assessment, 8, 220–234.

Clayson, D. E., Frost, T. F., & Shefflet, M. J. (2006). Grades and the student evaluation of Instruction: A test of the reciprocity effect. Academy of Management Learning & Education, 5, 52–65.

Costa, P. T., & McCrae, R. R. (1992). Revised NEO personality inventory and the NEO five-factor inventory. Odessa, FL: Psychological Assessment Resources.

De Dreu, C. K., Weingart, L. R., & Kwon, S. (2000). Influence of social motives on integrative negotiation: A meta-analytic review and test of two theories. Journal of Personality and Social Psychology, 78, 889–905.

Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. Behavioral and Brain Sciences, 22, 491–569.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: Ten aspects of the Big Five. Journal of Personality and Social Psychology, 93, 880–896.

Funder, D. C. (2012). Accurate personality judgment. Current Directions in Psychological Science, 21, 177–182.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), Handbook of personality psychology (pp. 795–824). San Diego, CA: Academic Press.

Guilford, J. P. (1954). Psychometric methods. New York, NY: McGraw-Hill.

Hagan, C. M., Konopaske, R., Bernardin, H. J., & Tyler, C. L. (2006). Predicting assessment center performance with 360-degree, top-down, and customer-based competency assessments. Human Resource Management, 45, 357–390.

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

Higgins, E. T. (1997). Beyond pleasure and pain. The American Psychologist, 52, 1280–1300.

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. Personnel Psychology, 63, 119–151.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? Psychological Methods, 5, 64–86.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 3, 424–453.

Jawahar, I. M. (2001). Attitudes, self-monitoring, and appraisal behaviors. Journal of Applied Psychology, 86, 875–883.

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

Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. Organizational Behavior and Human Decision Processes, 45, 194–208.

Kline, R. B. (1998). Principles and practice of structural equation modeling. New York, NY: Guilford Press.

Kneeland, N. (1929). That lenient tendency in rating. Personnel Journal, 7, 356–366.

Krishnan, S. K., & Singh, M. (2010). Outcomes in intention to quit of Indian IT professionals. Human Resource Management, 49, 421–437.

Le, H., Oh, I., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. Journal of Applied Psychology, 96, 113–133.

LeBlanc, P. W., & Mulvey, P. W. (1998). How American workers see the rewards of work. Compensation and Benefits Review, 30, 24–34.

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. Journal of Management, 30, 881–905.

Mero, N. P., Guidice, R. M., & Brownlee, A. L. (2007). Accountability in a performance appraisal context: The effect of audience and form of accounting on rater response and behavior. Journal of Management, 33, 223–252.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance evaluation: Social, organizational, and goal-based perspectives. Thousand Oaks: CA: Sage.

Ng, K., Koh, C., Ang, S., Kennedy, J. C., & Chan, K. (2011). Rating leniency and halo in multisource feedback ratings: Testing cultural assumptions of power distance and individualism-collectivism. Journal of Applied Psychology, 96, 1033–1044.

O'Neill, T., Goffin, R. D., & Gellatly, I. R. (2012). The use of random coefficient modeling for understanding and predicting job performance ratings: An application with field data. Organizational Research Methods, 15, 436–462.

Pichler, S. (2012). The social context of performance appraisal and appraisal reactions: A meta-analysis. Human Resource Management, 51, 709–732.

Pulakos, E. D., & O'Leary, R. S. (2011). Why is performance management broken? Industrial and Organizational Psychology, 4, 146–164.

Randall, R., & Sharples, D. (2012). The impact of rater agreeableness and rating context on the evaluation of poor performance. Journal of Occupational and Organizational Psychology, 85, 42–59.

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employee selection. Personnel Psychology, 50, 723–736.

Roch, S. G., Ayman, R., Newhouse, N., & Harris, M. (2005). Effect of identifiability, rating audience, and conscientiousness on rating level. International Journal of Selection and Assessment, 13, 53–62.

Rynes, S. L., Gerhart, B., & Minette, K. A. (2004). The importance of pay in employee motivation: Discrepancies between what people say and what they do. Human Resource Management, 43, 381–394.

Sackett, P., & Lievens, F. (2008). Personnel selection. Annual Review of Psychology, 59, 419–450.

Saucier, G. (1998). Replicable item-cluster subcomponents in the NEO Five-Factor Inventory. Journal of Personality Assessment, 70, 263–276.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. Psychological Review, 101, 632–652.

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

Sulsky, L., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 497–506.

Tyler, C. L., & Bernardin, H. J. (2003). A comparison of three managerial selection methods. Best Papers of the Academy of Management: HR, 2003. F 1-6. Paper presented at 2003 AOM meeting, Vancouver, BC, Canada.

Tziner, A., Murphy, K. R., & 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.

Villanova, P., Bernardin, H. J., Dahmus, S. A., & Sims, R. L. (1993). Rater leniency and performance appraisal discomfort. Educational and Psychological Measurement, 53, 789–799.

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.

Williams, M. L., McDaniel, M. A., & Nguyen, N. T. (2006). A meta-analysis of the antecedents and consequences of pay level satisfaction. Journal of Applied Psychology, 91, 392–413.

Wong, K. F. E., & Kwong, J. Y. Y. (2007). Effects of rater goals on rating patterns: Evidence from an experimental field study. Journal of Applied Psychology, 92, 577–585.

Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. Journal of Personality and Social Psychology, 99, 174–190.

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