

This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

The impact of rater agreeableness and rating context on the evaluation of poor performance

PLEASE CITE THE PUBLISHED VERSION

<http://dx.doi.org/10.1348/2044-8325.002002>

PUBLISHER

Wiley (© British Psychological Society)

VERSION

AM (Accepted Manuscript)

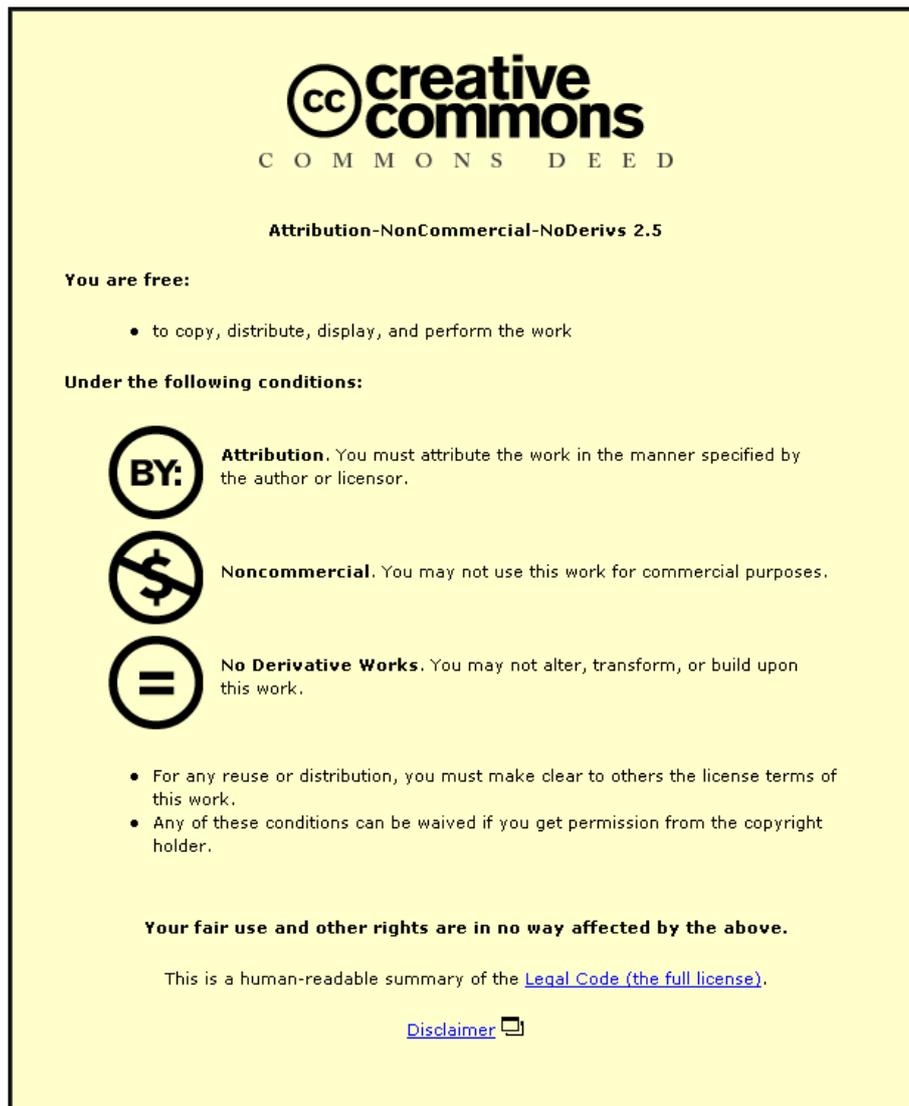
LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Randall, Raymond, and Daniel Sharples. 2019. "The Impact of Rater Agreeableness and Rating Context on the Evaluation of Poor Performance". figshare. <https://hdl.handle.net/2134/15223>.

This item was submitted to Loughborough's Institutional Repository (<https://dspace.lboro.ac.uk/>) by the author and is made available under the following Creative Commons Licence conditions.



CC creative commons
COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

BY: **Attribution.** You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the [Legal Code \(the full license\)](#).

[Disclaimer](#) 

For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

The impact of rater agreeableness and rating context on the evaluation of poor performance

Raymond Randall

School of Psychology, University of Leicester, UK

Daniel Sharples

Department for Work and Pensions, UK

Abstract

We tested the effects of rater agreeableness on the rating of others' poor performance in performance appraisal (PA). We also examined the interactions between rater agreeableness and two aspects of the rating context: ratee self-ratings and the prospect of future collaboration with the ratee after the feedback of PA ratings. Participants were government employees ($N = 230$) allocated to one of six experimental groups (a 3×2 between-groups design) or a control group ($n = 20$). Participants received accurate, low-deviated or high-deviated self-ratings from the ratee. Half were notified they would collaborate with the ratee in a future task. High rater agreeableness, positive deviations in self-rating, and the prospect of future collaboration were all independent predictors of higher PA ratings. The interactions between rater agreeableness and rating context were very small and inconsistent. We argue that conflict avoidance is an important motivation for those rating the performance of others.

Keywords: performance appraisal, agreeableness, self-rating, conflict

Leniency in subjective performance appraisal (PA) ratings has been found to be common and difficult to reduce (e.g., Kasten & Weintraub, 1999; Murphy & Cleveland, 1995; Murphy, Cleveland, Skattebo, & Kinney, 2004). It has been shown that raters are not always willing to rate accurately: Managers can make accurate private judgments of employee performance but that these may not be the same as the ratings they give in public (Murphy & Cleveland, 1995; Tziner, Murphy, & Cleveland, 2005). This could cause particularly serious problems when poor work performance needs to be identified and discussed (Arvey & Murphy, 1998; Bernardin, Tyler, & Villanova, 2009; Heslin, Latham, & VandeWalle, 2005).

Aspects of the rating context (e.g., the potential for conflict with the ratee) and rater disposition (e.g., high rater agreeableness and low conscientiousness) have all been linked to rater leniency (Tziner, Murphy, & Cleveland, 2005). However, the effects of the interactions between rating context and rater disposition on rating behaviour remain unclear. There has also been some debate about whether leniency is caused by rater disposition when very poor performance is rated (Yun, Donahue, Dudley, & McFarland, 2005; Bernardin, Tyler, & Villanova, 2009). In this study we used an experimental design within a working population to examine how rating context and rater agreeableness combined to influence PA ratings of poor performance.

Rater Agreeableness and Rating Context as Causes of Leniency in PA Ratings

Leniency in PA ratings has been linked to rater goals that run counter to the need for accuracy (e.g., Murphy et al., 2004; Murphy, Cleveland, Kinney, Skattebo, & Sin, 2004; Tziner, Latham, Price, & Haccoun, 1996). Lenient ratings may have contributed to empathic buffering (Waung & Highhouse, 1997) or helped raters to avoid the uncomfortable process of transmitting negative news (e.g., Blumberg, 1972). This type of rating behaviour may have

also been a product of a manager's fear of conflict with poor performing ratees or seen as a way to improve an employee's performance by increasing employees' self-efficacy (e.g., Longenecker, Sims, & Gioia, 1987). High-agreeableness individuals have shown a particular propensity to be motivated to give lenient ratings (Bernardin, Cooke, Ross & Villanova, 2000; Jawahar, 2001) perhaps because they value conflict-free, harmonious relationships with others and have a willingness to compromise their own interests for the sake of others (Goldberg, 1992). When combined with low rater conscientiousness, high agreeableness has been found to result in particularly lenient ratings (Bernardin, Tyler, & Villanova, 2009).

The main effects of rating context on rating behaviour have been well-established. A requirement for raters to provide feedback on their ratings has been linked to increased leniency (Murphy & Cleveland, 1995; Bernadin, Tyler, & Villanova, 2009). Kane, Bernardin, Villanova, and Peyrefitte (1995) and Harris (1994) have argued that the goal of the rater (e.g., to be lenient) could be a result of the rater's disposition but that this goal must be activated by the rating context. In the case of a high-agreeableness manager the context may activate strongly their concerns about harming their relationship with a subordinate (e.g., Bass, 1956; Murphy & Cleveland, 1995; Murphy et al., 2004). When poor employee performance was being evaluated high agreeableness managers may have been strongly motivated to use PA ratings to enhance employee self-efficacy or to avoid post-appraisal conflict and retaliatory decreases in employee performance (Harris, 1994).

Relatively little research has examined directly the interactions between rater disposition and rating context. Tziner, Murphy, Cleveland, Yavo, and Hayoon (2008) showed that raters' self-monitoring could interact with the rating context to influence rater behaviour. Yun, et al. (2005) found that the relationship between agreeableness and rating behaviour was strongest when the rating context (i.e. the need to provide face-to-face feedback) raised the possibility of conflict between the rater and the ratee. The same study cast some doubt on the

consistency of the effect of agreeableness of rating behaviour because agreeableness did not reliably predict rating leniency when poor performance was evaluated.

Yun et al. (2005) argued that an inconsistent relationship between agreeableness and PA ratings could have been the result of reduced potential for conflict when employee performance was poor. They suggested that high-agreeableness raters may have assumed that the poor performing ratee held an accurate view of their own performance (since it was bad enough to be self-evidently poor) and therefore the rater did not anticipate conflict. Other research has identified a consistent relationship between rater agreeableness and rating behaviour across all levels of ratee performance (Bernadin, Tyler, & Villanova, 2009). However, neither of the studies by Yun et al. (2005) or Bernadin, Tyler, and Villanova (2009) included a controlled manipulation of ratee self-ratings meaning that a direct test of the interaction between agreeableness and this aspect of rating context on the rating of poor performance was not conducted.

A main effect of self-ratings on others' PA ratings could occur because individuals tend to change their views toward the audience they are accountable to when the audience's views are known (Tetlock, 1983). Klimoski and Inks (1990) found that PA ratings were significantly higher when raters were made aware that the ratee had made a high self-rating. Shore, Adams and Tashchian (1998) showed that raters who received self-ratings highly positively deviated from actual performance were more inclined to elevate their own ratings. Murphy and Cleveland (1995) argued that such leniency tends to occur because managers want to protect subordinates' self-esteem or the manager-subordinate relationship. Therefore, it may be that agreeableness interacts with self-rating accuracy. High agreeableness raters may be more strongly motivated to give lenient ratings when they have knowledge that the ratee has over-estimated their own performance because this increases the potential for

conflict. Controlled testing of these interaction effects is important because the use of self-ratings in the PA process is increasingly common (CIPD, 2005; Hannum, 2007).

It may also be that the commonly used manipulation of a single discrete episode of face-to-face rating feedback does not introduce sufficient potential for conflict to motivate high-agreeableness raters to give lenient ratings. Actual manager-subordinate relationships usually involve the prospect of long-term co-dependency that goes beyond the provision of feedback (Ilgen & Favero, 1985). It is this co-dependent relationship that high agreeableness raters may be particularly keen to protect. There has been little research that has examined rater agreeableness alongside manipulations of the on-going relationship between the rater and the ratee beyond one discrete episode of PA rating feedback. Manipulations of on-going collaboration have been difficult to achieve but represent a more realistic representation of future contact (and hence the potential for conflict) within organizational PA process (Arvey & Murphy, 1998). Therefore, it remains unclear whether agreeableness interacts with on-going collaboration between the rater and the ratee to influence rating behaviour.

The Present Study

We conducted a controlled investigation of the main effects of rater agreeableness and rating context (and their interactions) on rating behaviour. The prospect of an on-going relationship between rater and ratee, and ratee self-ratings were manipulated. These independent variables were used an effort to control formally the potential for conflict between rater and ratee.

Examples of poor performance were used as rating stimuli in the study. This allowed us to test whether rater agreeableness was directly linked to the rating of poor performance. In doing so we examined whether the absence of PA rating leniency for high-agreeable raters assessing poor performance found by Yun et al. (2005) could be explained by insufficient potential for conflict in the rating context. The rating of poor performance is also a situation

in which the accurate measurement of, and feedback about, performance is particularly important in the organizational setting. In summary we tested three hypotheses.

Hypothesis 1: Anticipated future collaboration, higher positive deviations in ratee self-ratings and higher agreeableness will all be associated with higher ratings of poor performance.

Hypothesis 2: The interaction between rater agreeableness and anticipation of future collaboration with the ratee will be a significant predictor of rating behaviour.

Hypothesis 3: The interaction between rater agreeableness and the size of over-estimation of performance in ratee self-ratings will be a significant predictor of rating behaviour.

The study was also designed to improve on the ecological validity of previous controlled experimental investigations of rater agreeableness and rating behaviour by using a working population and the independent variables already described.

Method

Participants

Participants were 230 government employees. All were working at junior management grade or above and were responsible for administering welfare benefits. They were volunteers recruited through opportunity sampling. Participants were randomly allocated to one of six experimental groups (each containing 35 participants) or to a control group (20 participants). Experimental groups consisted of 108 females and 102 males with age ranging from 19 to 65 years ($M_{\text{age}} = 39.8$ years). The control group consisted of 12 females and 8 males with age ranging from 26 to 57 years ($M_{\text{age}} = 40.3$ years). The majority of the participants in the experimental groups were white ($n = 199$). Of the sample, 81% were in junior management grades, and the remaining participants in middle management grades. 201 of the participants in experimental groups were trained in the use of PA by their

organization, and 130 participants were currently responsible for conducting PAs for subordinates. Inspection of organizational records indicated that in terms of age, gender, ethnicity and grade the sample was representative of the population from which it was drawn.

Design

A 3×2 factorial design was used (see Table 1). The independent variables were self-rating deviation from accuracy (high, low or no deviation) and collaboration in a future task (anticipated or not anticipated). All participants in the six experimental groups were informed that they would be required to provide face-to-face feedback to a confederate. The three competency ratings given by each participant and their sum (a composite performance rating) formed the dependent variables.

The twenty participants in the control group were selected at random from the study population. Mean ratings from this group were used to test the significance of any leniency exhibited by the six experimental groups. The mean agreeableness scores and demographic profile of the control group did not differ significantly from that of any of the experimental groups. The advantages offered by using an experimental design meant that we could follow some but not all of Borman's (1978) recommendations for obtaining true PA ratings from control group participants. All those in the control group received training in order to provide some control over differences in previous rating experience; they were not exposed to the simulated organizational constraints that were introduced by the study; and they were not told that they would be required to provide face-to-face feedback to the ratee. All study participants were given the same access to a standardised set of observations of performance which were designed to map directly onto the rating scales. However, the design did restrict raters' access to observations about performance. In addition, the competencies used were broadly defined making behaviourally anchored rating scales unwieldy given the constraints on participants' time. Therefore, the control group score may not have reflected the true score

but provided a baseline against which the significance of the effects of the study variables could be established.

-Insert Table 1 about here-

The level of control exerted over performance data presented to the rater and the on-going rater-ratee relationship could not have been achieved in a functioning organisational PA system because of significant practical and ethical constraints. For example, the relationship between employees could be irrevocably damaged if self-ratings of performance were manipulated and the random allocation of participants to the experimental conditions would not be possible.

Measures

Demographic data. Participants completed a seven-item demographics questionnaire to assess participant grade, previous PA training, current responsibility for conducting PAs, how many PAs conducted annually, gender, age and ethnicity. The latter three have been identified as factors that can influence ratings (Furnham & Stringfield, 2001; Lefkowitz, 2000; Schwab, & Heneman, 1978).

Agreeableness measure. Participant agreeableness was measured using the 20 agreeableness items from Goldberg's (2001) International Personality Inventory Pool (IPIP). Example items include "I am easy to satisfy" and "I suspect hidden motives in others". Participants responded to the items using a five-point scale (from 1 = *very inaccurate* to 5 = *very accurate*). The inventory produced a single overall score of agreeableness for each participant (minimum = 20, maximum = 100). The IPIP-NEO agreeableness items are independent of the other Big Five domains (Goldberg, 1992) and in this study demonstrated acceptable internal reliability (Cronbach's $\alpha = .80$).

Performance rating. Participants were provided with a written report of the performance of the ratee and notified that the researcher wrote the report from observations

taken. This was developed from a situational interview question from Arnold et al. (2005, p. 181). Three variants of the report were written and the one most representative of poor performance was selected by a focus group of four human resources managers. The experimental scenario consisted of three parts: background to the event being assessed, a qualitative report of the ratee's performance during the situation (described in line with the three competencies), and instructions on how to use the rating scale.

When making their ratings, participants used a single item nine-point Likert scale (from 1 = *poor performance* to 9 = *excellent performance*) for each of three competencies (each illustrated with example behaviours). These were: initiative (described to participants as “demonstrating an ability to actively start something new, to independently develop novel solutions to existing problems and to actively engage with these problems”); maximizing profitability (described as “demonstrating an ability to take decisions that enable the organization to maximize profits and maintain its position within the market”); and ensuring the well-being of employees (described as “demonstrating an ability to take decisions that ensure the welfare of employees is taken into consideration and that they are continually protected within the workplace”). These three competencies were chosen because they were similar to the competencies used in the participants' organisation and may resemble competencies used by various organisations.

Before giving their ratings, participants received one-hour of rater training. This included: a description of the background to the rating task; presentation of the competency framework; and instructions on assessing performance using the rating scale and self-ratings. Participants then completed two practice rating exercises that replicated the experimental task. This training was a shorter (but not narrower) version of the PA training provided by the organisation (i.e. briefer coverage of the same training content including: how to use rating scales, how to make evaluations from performance data and how to rate accurately). This

training was used because previous research has shown that providing rater training can reduce the risk of participants over-relying on ratee self-rating (Shore & Tashchian, 2002).

Manipulations

Face-to-face feedback. All participants in the six experimental groups were told to anticipate face-to-face feedback in order to elicit the link between rater personality and rating behaviour (Yun et al., 2005). A confederate (24-year old white male dressed in smart work attire, who was instructed to smile and acknowledge the participants but did not speak to them) was introduced to participants by the experimenter using the phrase “this is the employee that you will be rating and giving face-to-face feedback to”. This was done during their participants’ training to strengthen the manipulations.

Self-rating reports. The reports from the ratee were either accurate (confirming poor performance across all three competencies), low-deviated (stating that performance was satisfactory across all competencies), or high-deviated self-ratings (stating that performance was excellent across all competencies). The material used for each of the three levels of this variable was chosen by the focus group of four human resources managers. The self-ratings consisted of three sections: an explanation of the situation by the ratee; the ratee’s qualitative assessment of their performance in reference to the three competencies; the ratee’s own rating on Likert-type scales for each competency (*1 = poor performance to 9 = excellent performance*). These ratings represented either accurate self-ratings (Initiative rating = 1, Profitability rating = 2, Well-being rating = 2), low deviated self-ratings (Initiative = 5, Profitability = 5, Well-being = 5) or high-deviated self-ratings (Initiative = 8, Profitability = 7, Well-being = 8). The ratings were determined by the focus group through their analysis of the qualitative data in the fabricated ratees’ accounts.

Future collaboration. Participants in the future collaboration conditions were notified by the researcher that they would be required to collaborate, in person, with the ratee

at some point in the future (after a face-to-face feedback session). The experimenter used the phrase “after feedback you will work with this person again as part of the development of the company appraisal system” (see also Experimental Procedure). No information about the prospect of future collaboration after the face-to-face feedback session was given to the other groups.

Manipulation checks. All participants in the experimental groups answered two written questions to assess effectiveness of the face-to-face feedback manipulation. These were: “Will you now be required to provide face-to-face feedback to the employee?” and “Will you now be required to work in a future task with the employee?” Following debriefing, participants were also asked, face-to-face, whether they believed the manipulations. Half received a positively phrased question: “Did you expect that you would have to provide face-to-face feedback to the employee?” The other half were asked a negatively phrased question: “Did you expect that you would not have to collaborate with the employee in a future task?”

Experimental Procedure

Data was collected in same-condition sessions with between 10 and 15 participants. Participants were informed that they were testing an appraisal system developed for a small local manufacturing company and that they would rate the performance of one of the managers from this company using a competency framework. Participants within the six experimental conditions were told by the researcher that they would be required to provide face-to-face feedback individually to the ratee. Participants within the future collaboration conditions were also instructed by the researcher that they would be required to work with the ratee on a task some time after providing feedback.

Participants then received training, completed two practice rating exercises, and were introduced to, but did not interact with, the ratee (the confederate) who vacated the room after

the being introduced. In the same format as the training exercises, participants were given written observations of the ratee's performance (just over half a page of A4 typescript) and then one of three different self-rating reports (high-deviation, low-deviation, or accurate). Participants then rated performance.

Participants were then asked to complete the agreeableness inventory and were assured of confidentiality to reduce unwanted trait variance (Jackson, Ashton, & Tomes, 1996). Participants in the six experimental conditions then completed the written manipulation check questions. All participants then provided demographic information. Participants were then debriefed and asked (verbally) whether they had believed that they believed the manipulations. Finally, they were requested not to discuss the details of the experiment with colleagues.

Data Analysis

Differences between the competency ratings given by each of the six experimental groups and the control group were explored using a separate ANOVA analyses for each competency rating. Tukey's HSD post-hoc tests were used to identify the significant differences between each of the intervention groups and the control group.

Analysis of covariance was then carried out for each dependent variable to test for the main effects and interaction effects of the independent variables and rater agreeableness, while controlling for demographic factors. Interaction terms were calculated using centered variables to avoid problems with multicollinearity¹.

A further set of analyses was used to identify the practical significance of the findings. Each of the six experimental group was divided into three sub-groups (low (>1 *SD* below the mean), high (>1 *SD* above the mean), and medium agreeableness). To test for significant rating leniency the mean performance rating (produced from a sum of the three

¹ In a regression model of the data the maximum VIF value using the centered variables to calculate interaction terms was 1.21. The same analysis with non-centered variables resulted in several VIF figures in excess of 10.

competency scores) of each of the 18 resultant groups was compared to the control group rating using independent samples t-tests² (with an adjusted level of significance of $p < .002$ applied). This analysis identified the circumstances under which the effects of the two independent variables and rater agreeableness worked together to produce ratings that were significantly different from the mean control group rating.

Results

Equivalence of Experimental Groups

The random allocation of participants produced six groups that were equivalent in terms of mean age, $F(5, 201) = .60, p = .70$, and gender composition (maximum $\chi^2(1, N = 70) = 0.92, p = .34$). A one-way ANOVA showed that mean scores and variances on the agreeableness trait were not significantly different across all six experimental conditions, $F(5, 204) = .60, p = .70$; Levene's statistic = .29, $p = .90$. All six experimental groups reported agreeableness scores that were not significantly different to those of the control group. The mean for the agreeableness measure ($M = 61.14$) was slightly above the scale mid-point of 60. Group six (anticipated collaboration and low self-rating deviation) had the highest number of non-white participants (four), whilst group five (anticipated collaboration and accurate self-rating) had no non-white participants (see Table 3). Those currently carrying out performance appraisals as part of their own job were over-represented in the experimental groups exposed to higher degrees of self-rating deviation. Therefore this demographic variable was included as a covariate in subsequent ANCOVA analyses.

Manipulation Checks

The experimental manipulations appeared successful. The proportion anticipating face-to-face feedback was consistently high across the experimental groups that were

² The Bonferroni correction was applied to significance testing i.e. the p -value was set at $p < .002$

exposed to this manipulation. It was 97% to 100%³ with the questionnaire, and 94% to 100% with the verbal enquiry. The proportion anticipating future collaboration was also high in the groups exposed to this manipulation. It was between 94% and 100% with the questionnaire, and between 91% and 97% with the verbal enquiry.

Descriptive Statistics and Correlations between Variables

All independent and dependent variables were approximately normally distributed. Across the sample as a whole, t-tests showed that ratings on the profitability scale were significantly higher than on both the well-being scale, $t(209) = 2.57, p < .05$, and the initiative scale, $t(209) = -8.58, p < .001$. Ratings on the well-being scale were also significantly higher than ratings on the initiative scale, $t(209) = -9.19, p < .001$. There was preliminary evidence for significant main effects for both independent variables and agreeableness: self-rating deviation, future collaboration and agreeableness were all positively correlated with the three dependent variables (see Table 2). However, agreeableness was not significantly correlated with ratings in the control group: This indicated that face-to-face feedback was a necessary condition to establish a relationship between rater agreeableness and PA ratings. Those responsible for PA in their current job role also tended to produce higher ratings for the three competencies ($ps < .05$). However, these correlations were small, $rs(208) < .17$, and appeared to be due to these participants being over represented in the experimental groups where rater self-ratings were highly deviated, $r(208) = .38, p < .01$. Age showed a small, positive correlation with ratings on the profitability competency, $r(208) = .15, p < .05$, suggesting that older raters tended to give higher ratings than younger raters. Significant positive correlations were found between the three competency ratings: The highest was $r(208) = .64$ (between profitability and initiative). All three competencies were treated as separate dependent variables. The ANCOVA analysis was also repeated using a composite measure of the

³ Of these two figures, the first is for the positively phrased question, the second for the negatively phrased question.

competencies (i.e. the sum of the three competencies ratings) as the dependent variable. The composite measure had good internal consistency (Cronbach's $\alpha = .82$). This was done to reflect the common organisational practice of summing competency ratings and to provide some protection against the problems with reliability that can occur with the use of single-item measures.

-Insert Table 2 about here-

Hypothesis Testing

Table 3 shows that significantly higher ratings were given by raters in the experimental groups when compared to raters in the control groups when there was: i) anticipated collaboration and some degree of positive deviation in self-ratings for all three competency ratings (Table 3, groups six and seven) and ii) no anticipated future collaboration and a high positive deviation in self-rating, but only for the well-being competency rating (see Table 3, group four).

-Insert Table 3 about here-

These findings indicate that raters did not generally give lenient ratings when exposed to inflated self-ratings if future collaboration was not anticipated. Where future collaboration was anticipated, ratings were significantly higher when raters were made aware that the ratees had over-estimated their performance. Only when ratee self-rating was accurate did the prospect of future collaboration not result in lenient ratings.

The ANCOVA analysis (Table 4) showed good explanatory power and indicated that the group differences were driven by the effects of several variables. 46% of the variance in initiative ratings and 47% of the variance in profitability ratings was accounted for by the sum of the effects of the independent variables, rater agreeableness and a very small three-way interaction effect. Together rater agreeableness and rating context accounted for a total of 37% of the variance in well-being ratings.

Across all three individual competency PA ratings significant main effects of rater agreeableness, self-rating deviation and anticipated future rater-ratee collaboration were identified (see Table 4). Higher levels of agreeableness, self-rating deviation and anticipated future collaboration predicted higher performance ratings. There was strong support for hypothesis 1. There were large effect sizes for agreeableness, $F(1, 193) = 79.18, p < .001, \eta_p^2 = .29$, and self-rating deviation, $F(2, 193) = 28.77, p < .001, \eta_p^2 = .23$, on profitability ratings, with a more modest effect being found for future collaboration, $F(1, 193) = 30.30, p < .001, \eta_p^2 = .14$. A similar pattern of results was found for initiative ratings, with large effects for agreeableness, $F(1, 193) = 51.84, p < .001, \eta_p^2 = .21$, and self-rating deviation, $F(2, 193) = 25.99, p < .001, \eta_p^2 = .21$, and a more modest effect being found for future collaboration, $F(1, 193) = 37.02, p < .001, \eta_p^2 = .16$. With well-being as the dependent variable, agreeableness had a large effect on performance ratings, $F(1, 193) = 70.95, p < .001, \eta_p^2 = .27$, but self-rating deviation, $F(2, 193) = 14.83, p < .001, \eta_p^2 = .13$, and future collaboration $F(1, 193) = 24.28, p < .001, \eta_p^2 = .11$, had more modest effects on ratings.

There was very limited support for hypotheses 2 and 3, with there being no significant two-way interactions. There were significant three-way interactions that explained variance in the dependent variables profitability and initiative. However, effect sizes were small, $F(1, 193) = 4.20, p < .05, \eta_p^2 = .02$, and $F(1, 193) = 5.81, p < .05, \eta_p^2 = .03$ respectively, in comparison to the size of the main effects. The absence of large interaction effects indicated that rater agreeableness, ratee self-ratings and anticipated future collaboration all exerted relatively independent effects on participants' ratings.

-Insert Table 4 about here-

The ANCOVA analysis with a composite performance measure as the dependent variable produced very similar findings to the ANCOVAs carried out for the three separate competencies. Significant main effects were found for rater agreeableness, $F(1, 193) =$

124.11, $p < .001$, $\eta_p^2 = .39$, future collaboration, $F(1, 193) = 56.42$, $p < .001$, $\eta_p^2 = .23$, and self-rating deviation, $F(2, 193) = 42.05$, $p < .001$, $\eta_p^2 = .30$. Together these variables explained 58% of the ratings in the composite measure. There were no other significant effects.

Figure 1 (showing results from the no future collaboration conditions) and Figure 2 (showing results from the future collaboration conditions) highlight the conditions under which PA ratings were significantly different to those given by the control group. It was apparent that a baseline level of rating leniency was determined, at least in part, by rater agreeableness. Therefore, the eventual outcome (i.e. the practical significance) of adding the effects of the rating context was different at each level of agreeableness. For example, for low-agreeableness raters the impact of the rating context did not lead to PA ratings that were significantly higher than those given by the control group in any of the experimental conditions.

-Insert Figure 1 about here-

- Insert Figure 2 about here -

In only one condition (no anticipated collaboration and no deviation in ratee self-rating) did high agreeableness raters produce a rating on the composite performance rating that was not significantly different to those of the control group. In conditions where future collaboration was not anticipated (see Figure 1), some over-estimate of performance by the ratee was all that was needed for high-agreeableness raters to produce ratings that were significantly higher than those given by the control group. When future collaboration was anticipated (see Figure 2) no self-rating deviation was needed for high-agreeableness raters to produce such significantly elevated ratings.

A more mixed picture emerged for medium-agreeableness raters who produced significantly higher ratings than the control group when there was anticipated collaboration

and some degree of positive self-rating deviation (see Figure 2). Without anticipated future collaboration these raters only produced significantly higher ratings than the control group they were aware that the ratee had made a large over-estimate of their own performance (see Figure 1).

An apparent divergence of the medium- and high-agreeableness plot lines in Figure 1 indicated that an Agreeableness \times Self-rating Deviation interaction effect have occurred across some but not all experimental groups. To test this, the ANCOVA analysis used to test the study hypotheses was repeated but using data from the participants in the accurate and low-deviation self-rating conditions where there was no prospect of future collaboration. This revealed a small but significant Agreeableness \times Self-rating Deviation interaction effect, $F(1, 61) = 4.19, p < .05, \eta_p^2 = .06$, and two other significant effects: a large main effect for agreeableness, $F(1, 61) = 31.63, p < .001, \eta_p^2 = .34$, and a significant medium-sized effect for deviated self-ratings, $F(1, 61) = 11.13, p < .001, \eta_p^2 = .15$. The model explained 46% of the variance in the composite PA rating. Independent samples t-tests showed that this interaction effect was the result of a significant difference in the composite PA rating between the mid-agreeableness group ($M = 6.00, SD = 0.88$) and high agreeableness group ($M = 7.60, SD = 1.57$) in the low self-rating deviation condition, $t(27) = -2.97, p < .05$. This difference was not present in the accurate self-rating condition.

Discussion

There has been debate about the whether the effect of rater agreeableness on PA ratings of poor performance is moderated by the rating context. Our results show that rater agreeableness has a strong effect on the ratings of poor performance the size of which altered very little by the manipulation of the rating context. We found that significant leniency in PA ratings of poor performance was the product of the sum of the independent effects of

positively deviated ratee self-ratings, the prospect of future collaboration and rater agreeableness.

In contrast to Yun et al. (2005) we found that even when rating poor performance, rater agreeableness exerts a largely independent effect on rating behaviour: The higher the rater's agreeableness the higher the ratings that were given. In line with Klimoski and Inks (1990) the results show that making highly deviated self-ratings available to raters produces the highest ratings. However, it extends their findings by showing that the impact of self-ratings are to a large extent independent of rater agreeableness and the prospect of future collaboration. What caused rating behaviour to be significantly different from that of a control group was the cumulative impact of unrelated dispositional (i.e. agreeableness) and contextual (i.e. the prospect of on-going collaboration and ratee self-ratings) factors. As Table 3 and Figure 1 show, none of the main effects were sufficient on their own to induce performance ratings significantly above those of the control group. The implication of this is that high agreeableness raters are more likely to produce lenient ratings in various rating contexts (see Figures 1 and 2). This is not because rating context strengthens the link between agreeableness and rating behaviour because high agreeableness leads to an elevated baseline level of PA ratings. Agreeableness provides an underlying motivation to provide lenient ratings that means that the additional effects of context are more likely to result in significant leniency for high-agreeableness raters.

In conditions that minimize the possibility of future conflict high agreeableness was the only stimulus for significant rating inflation (see Figure 1) but was not sufficient to produce ratings that were significantly higher than those made by the control group (even with the prospect of face-to-face feedback). At first this may look like the type of interaction effect (i.e. that the rating context stimulates lenient ratings by high agreeableness raters) that was suggested by Kane et al. (1995). However, as Figures 1 and 2 show, this result is

indicative of the separate, but additive effects that disposition and context have on rater behaviour. In the same low potential for conflict conditions these effects may mean that low-agreeableness raters provide negative PA ratings (see Figure 1).

There was little support for hypotheses 2 and 3. Only very small three-way interaction effects were present for two of the competencies. The only exception to this was for the composite measure of performance in conditions where there was no anticipated future collaboration where rater agreeableness amplified the effects of introducing positively deviated ratee self-ratings. Our results indicate that high agreeableness raters are particularly susceptible to influence by small inaccuracies in others' self-ratings when there is no prospect of future collaboration beyond the provision of face-to-face feedback. This may explain why Yun et al. (2005) failed to find a link between rater agreeableness and the rating of poor performance in the absence of a manipulation of ratee self-ratings. Further research should also investigate whether a critical mass of three co-occurring conditions (inflated self-ratings, future collaboration and high rater agreeableness) leads to greater rating leniency than can be explained by adding together the independent effects of these factors.

When there is some potential for conflict in the rating context low-agreeableness raters provide higher PA ratings that were similar to those of the control group (see Figures 1 and 2). The organizational PA context is likely to often provide such potential for conflict. It is common for people to over-estimate their performance (Dunning, 2006) and for those who conduct the appraisal to continue to work in collaboration with the employee in the future. Therefore we suggest that raters who are low on agreeableness may provide the most accurate PA ratings of poor performance in organizational settings.

Our results indicate that conflict avoidance was the most likely explanation for rating leniency. The results are inconsistent with the hypotheses that leniency is the result of a desire to improve ratee self-efficacy or is a result of a general tendency to avoid negative

feedback. If these were the dominant motives, lenient ratings would be given without the potential for conflict introduced through the manipulations of context carried out in this study. However, raters were not directly asked about their motives. By giving generous ratings it may be that high-agreeableness raters are concerned about the ratee's long-term view of the rater and focused on developing a harmonious working relationship. Low-agreeableness raters may view the provision of realistic feedback as central to the goal of improving employee performance. Our results also suggest that potential for conflict in the rating context might blunt the effects of low agreeableness raters' tendency to provide negative Pas (see Figures 1 and 2).

Implications for Identifying Poor Performance in Organizations

We suggest that rater training should include a component designed to help managers manage the potential for conflict that appears likely in many organizational PA contexts. It has already been shown that self-efficacy can also have a significant positive impact on the accuracy of rating behaviour (Bernardin & Villanova, 2005; Tziner, Murphy & Cleveland, 2005). Therefore PA training could include components specifically designed to help all raters develop self-efficacy for dealing with conflict. Managers of various levels of agreeableness may not be aware that inflated ratee self-ratings can lead them to make lenient ratings. Using training to make managers aware of this effect may help to reduce leniency particularly if managers are also informed that PA rating leniency is not an appropriate way to deal with poor performance. Further research is needed to investigate the impact of training interventions designed to reduce the strength of the main effects found in this study. Rater training could be combined with interventions that make raters accountable for their ratings (Mero, Guidice & Brownlee, 2007) and to reward raters for the accuracy of their ratings (Murphy & Cleveland, 1995). It may also be that PA ratings need to be collected from low agreeableness raters in work situations where the accurate identification of poor

performance is especially important (e.g., when poor employee performance places safety at risk).

Our results indicate that modifications to the rating context can reduce leniency. When accurate PA ratings are especially important managers might need to be denied access to employees' self-ratings until they have made their own PA ratings. DeGregorie and Fisher (1988) have found that this intervention can cause employee dissatisfaction with the PA process. Those involved in the design and implementation of rating processes will therefore need to make informed decisions about whether accurate ratings or employee satisfaction with the process are of greater importance. The issue of future collaboration with the ratee is a more difficult avenue of intervention. We found that the effect of this variable on PA ratings was significant but modest in comparison to the effects of rater agreeableness and making available inaccurate self-ratings. It may be that the prospect of future collaboration introduces an unavoidable baseline level of PA leniency. This may not a significant problem if other aspects of the rating context have been designed to reduce the impact of other sources of leniency (see Table 3 Group 5) and rater training proves effective in reducing the effects of agreeableness on PA ratings.

Limitations

Experimental PA research has often been criticised for its failure to capture the social milieu within which organizational raters are situated (Dipboye, 1990). Although this study did not include a complete simulation of the work environment, we used a working population and controlled manipulations of social pressures to achieve some gains in ecological validity over much previous research on PA rating behaviour. The effect size for the prospect of future collaboration may have been larger if participants were given more detailed information about the collaboration to make it more realistic (e.g., its timing and a more precise description of the nature of the collaborative task). The use of written reports

has the advantage of enhancing control the rater's exposure to the evidence upon which ratings were made: this would not be possible in a functioning organisation. Some researchers consider it inappropriate to use textual information as the basis for PA ratings (e.g., Jawahar & Williams, 1997). In an attempt to bolster ecological validity we used a carefully constructed procedure that allowed participants to meet the person whose performance was being rated. Also given the increased use of sophisticated information and communication technologies, it is increasingly common for and for textual information (e.g., emails) to be a significant source of information about problems with employee performance especially when subordinates work remotely from their manager (Rockart, 1998).

In our study raters did not have any prior knowledge of the ratee. Such knowledge could have a significant affect on rating behavior. For example, research has identified that the degree of similarity between rater and ratee can impact on ratings (Cardy & Dobbins, 1986). Conditions in this study do not account for this, yet there would be significant ethical problems associated with carrying out controlled manipulations of the rating context around established employee-manger relationships. Random allocation of participants was used to protect against the impact of a range of unmeasured variables that may impact on rating behaviour, such as participants own recent appraisal results (Latham, Budworth, Yanar, & Whyte, 2008). The use of participants from the same organization could mean that the results were influenced by factors such as the organizational culture that exert a unique influence over study population (Tziner, Murphy, & Cleveland, 2005). Further research is needed to examine whether rater agreeableness interacts with participants own experiences of PA process to determine rating behaviour and to test whether our results are replicated among employees from other organizations.

Rater conscientiousness is likely to explain additional variance in rating behaviour. This rater disposition was not included in this study because our aim was to resolve debates

about the interactions between rating context and rater agreeableness in the rating of poor performance. Any effects of rater conscientiousness are likely to be independent of those found in this study: Agreeableness and conscientiousness are orthogonal factors. Moreover, the introduction of the potential for conflict would not be expected to interact with rater conscientiousness because the motivational mechanisms thought to underpin the PA rating behaviour of highly conscientious individuals are not linked to conflict avoidance (Bernardin et al., 2000; Yun et al., 2005). Future research should investigate the cumulative effects of conscientiousness and agreeableness on rating behaviour in various rating situations (e.g., when there is the potential for conflict or where accurate rating is rewarded).

Conclusion

There are significant independent effects of rater agreeableness and rating context on PA ratings of poor performance. These need to be carefully assessed and managed. Of the variables studied, rater agreeableness had the largest effect on PA ratings and the prospect of future collaboration had the smallest effect. All three effects appear to be driven by raters' motivation to avoid conflict with the ratee. The interactions between these variables had a very small and inconsistent effect on rating behaviour. Research is needed to test whether a reduction in rater leniency occurs through interventions such as the provision of rater training that helps raters handle conflict situations and restricting raters' access to ratee self-ratings.

References

- Arnold, J., Silvester, J., Patterson, F., Robertson, R., Cooper, C., & Burnes, B. (2005). *Work psychology: Understanding human behaviour in the workplace* (4th ed.). Harlow: Prentice Hall.
- Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology*, *49*, 141-168.
- Bass, B. M. (1956). Reducing leniency in merit ratings. *Personnel Psychology*, *9*, 359-369.
- Bernardin, H. J., Cooke, D. Ross, S., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology*, *85*, 232-236.
- 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.
- Bernardin, H. J., & Villanova, P. (2005). Research streams in rater self-efficacy. *Group and Organization Management*, *30*, 61-88.
- Blumberg, H. H. (1972). Communication of interpersonal evaluations. *Journal of Personality and Social Psychology*, *23*, 157-162.
- Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology*, *63*, 135-144.
- Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. *Journal of Applied Psychology*, *71*, 672-678.
- Chartered Institute of Personnel and Development (2005). Survey report September 2005: Performance management. London: Chartered Institute of Personnel and Development.

- DeGregorie, M., & Fisher, C. D. (1988). Providing performance feedback: Reactions to alternative methods. *Journal of Management, 14*, 605-616.
- Dipboye, R. L. (1990). Laboratory vs. field research in industrial and organizational psychology. *International Review of Industrial and Organizational Psychology, 5*, 1–34.
- Dunning, D. (2006). Strangers to ourselves? *The Psychologist, 19*, 600-603.
- Furnham, A., & Stringfield, P. (2001). Gender difference in rating reports: Female managers are harsher raters, particularly of males. *Journal of Managerial Psychology, 16*, 281-289.
- Goldberg, L. R. (1992). The development of markers for the Big 5 factor structure. *Psychological Assessment, 4*, 26-42.
- Goldberg, L. R. (2001). Possible questionnaire format for administering the 100-item set of IPIP Big-Five Factor markers. Retrieved 8th December, 2005 from http://ipip.ori.org/New_IPIP-100-item-scale.htm.
- Hannum, K. M. (2007). Measurement equivalence of 360° assessment data: Are different raters rating the same constructs? *International Journal of Selection and Assessment, 15*, 293-301.
- Harris, M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*, 737-756.
- Heslin, P. A., Latham, G. P., & VandeWalle, D. (2005). The effect of implicit person theory on performance appraisals. *Journal of Applied Psychology, 90*, 842-856.
- Ilgén, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review, 10*, 311-321.

Jackson, D. N., Ashton, M. C., & Tomes, J. L. (1996). The six-factor model of personality: Facets from the Big-Five. *Personality and Individual Differences, 21*, 391-402.

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

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*, 905-926.

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

Kasten, R., & Weintraub, Z. (1999). Rating errors and rating accuracy: A field experiment. *Human Performance, 12*, 137-153.

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

Latham, G. P., Budworth, M-H., Yanar, B., & Whyte, G. (2008). The influence of a manager's own performance appraisal on the evaluation of others. *International Journal of Selection and Assessment, 16*, 220-228.

Lefkowitz, J. (2000). The role of interpersonal affective regard in supervisory performance ratings: A literature review and proposed causal model. *Journal of Occupational and Organizational Psychology, 73*, 67-86.

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1*, 183-193.

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 appraisal: Social, Organizational and goal-based perspectives*. Thousand Oaks, CA: Sage Publications.
- Murphy, K. B., Cleveland, J. N., Kinney, T. B., Skattebo, A. L., Newman, D. A., & Sin, H. P. (2004). Unit climate, rater goals and performance ratings in an instructional setting. *Journal of Management*, 24, 48-65.
- Murphy, K., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology*, 89, 158-164.
- Rockart, J. (1998). Towards survivability of communication-intensive new organization forms. *Journal of Management Studies*, 35, 417-420.
- Schwab, D. P., & Heneman, H.G. (1978). Age stereotyping in performance appraisal. *Journal of Applied Psychology*, 63, 573-578.
- Shore, T. H., Adams, J. S., & Tashchian, A. (1998). Effects of self-appraisal information, appraisal purpose, and feedback on performance appraisal ratings. *Journal of Business and Psychology*, 12, 283-298.
- Shore, T. H., & Tashchian, A. (2002). Accountability forces in performance appraisal: Effects of self-appraisal information, normative information and task performance. *Journal of Business and Psychology*, 17, 261-274.
- Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 52, 74-83.
- Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and rater factors affecting rating behavior. *Group and Organization Management*, 30, 89-98.

Tziner, A., Latham, G. P. Price, B. S., & Haccoun, R. (1996). Development and validation of a questionnaire for measuring perceived political considerations in performance appraisal. *Journal of Organizational Behavior, 17*, 179-190.

Tziner, A., Murphy, K. R., Cleveland, J. N., Yavo, A. & Hayoon, E. (2008). A new old question: Do contextual factors relate to rating behavior: An investigation with peer evaluations. *International Journal of Selection and Assessment, 16*, 59-67.

Waung, M., & Highhouse, S. (1997). Fear of conflict and empathic buffering: Two explanations for the inflation of performance feedback. *Organizational Behavior and Human Decision Processes, 71*, 37-54.

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.

Table 1

Experimental Design

Future collaboration (two levels)		
	Anticipated collaboration	Not anticipated collaboration
Ratee self-rating (three levels)	High positive deviation from accurate ratee self-rating	High positive deviation from accurate ratee self-rating
	Low positive deviation from accurate ratee self-rating	Low positive deviation from accurate ratee self-rating
	No deviation from accurate ratee self –rating (accurate)	No deviation from accurate ratee self –rating (accurate)

Table 2

Means, Standard Deviations and Correlations (Experimental Groups Only)

Variable	M	SD	1	2	3	4	5	6	7	8	9	10
1. Age	39.76	10.85	-									
2. Gender ^a	1.49	.50	-.02	-								
3. Ethnicity ^b	1.05	.22	-.04	.07	-							
4. Current use of PA ^c	1.41	.49	-.08	-.03	.02	-						
5. Grade ^d	1.24	.53	.27**	.03	.01	.12	-					
6. Self-rating deviation ^e	2.00	.82	.08	-.03	.08	.38**	.08	-				
7. Collaboration ^f	1.50	.50	.05	.20	-.02	-.02	.05	.00	-			
8. Agreeableness	61.14	17.03	.01	.04	.00	.00	.00	.07	-.11	-		
9. Initiative	2.35	.84	.08	-.01	.03	.15*	.12	.45**	.29**	.37**	-	
10. Profitability	2.35	.90	.15*	-.05	.10	.15*	.03	.45**	.24**	.49**	.64**	-
11. Wellbeing	2.01	.80	.03	.00	.19**	.16*	-.01	.36**	.23**	.41**	.58**	.58**

N = 210. * $p < 0.05$; ** $p < 0.01$ (two tailed)

Note. ^a Gender coded as 1 = female; 2 = male. ^b Ethnicity coded as: 1 = white participants; 2 = non-white participants. ^c Current use of PA coded as 1 = yes; 2 = no. ^d Grade coded as 1 = junior managers; 2 = middle managers (low grade); 3 = middle managers (high grade). ^e Self-rating deviation coded as 1 = no deviation; 2 = low deviation; 3 = high deviation. ^f Collaboration coded as 1 = no anticipated collaboration; 2 = anticipated collaboration.

Table 3

Mean Competency Rating Scores for the Six Experimental Conditions and the Control Group

Group	Ratee self-rating	Future collaboration	Male	Female	Ethnicity ^a	Mean age (SD)	Mean Agreeableness (SD)	Mean Initiative Rating (SD)	Mean Profitability Rating (SD)	Mean Well being Rating (SD)
1.	Control group	Control group	8	12	3	40.35 (9.50)	63.05 (11.51)	1.95 (.69)	1.75 (.61)	1.55 (.60)
2.	Accurate	Not anticipated	19	16	1	37.89 (8.20)	60.26 (17.42)	1.69 (.68)	1.57 (.61)	1.49 (.51)
3.	Low positive deviation	Not anticipated	15	20	2	39.89 (13.66)	63.97 (17.15)	2.09 (.74)	2.26 (.74)	1.80 (.58)
4.	High positive deviation	Not anticipated	16	19	3	39.89 (12.64)	63.37 (16.54)	2.54 (.78)	2.57 (.92)	2.20* (.68)
5.	Accurate	Anticipated	17	18	0	38.39 (9.03)	59.91 (18.69)	2.06 (.64)	2.00 (.64)	1.74 (.66)
6.	Low positive deviation	Anticipated	18	17	4	41.73 (9.06)	58.00 (16.26)	2.69* (.68)	2.71* (.75)	2.40* (.85)
7.	High positive deviation	Anticipated	17	18	1	40.77 (11.51)	61.34 (16.49)	3.06* (.76)	2.97* (.95)	2.43* (.95)

Note. * indicates mean ratings significantly different from the control group at $p < .008$ (Bonferroni correction applied). The maximum between-condition ratios of variance in the dependent variables were 1.49 for initiative ratings, 2.46 for profitability ratings and 3.50 for well-being ratings.

^a Numbers given indicate number of non-whites in the group.

Table 4

ANCOVA Results for Tests of the Effects of Self-rating Deviation, Rater Agreeableness and Anticipated Collaboration on Competency Ratings.

Predictor variables	df	Dependent variable: Initiative rating			Dependent variable: Profitability rating			Dependent variable: Well-being rating			
		MS	F	η_p^2	MS	F	η_p^2	MS	F	η_p^2	
Gender	1	.07	.17	.00	.49	1.14	.01	.07	.19	.00	
Age	1	.00	.00	.00	1.43	3.31	.02	.00	.03	.00	
Ethnicity	1	.00	.00	.00	.60	1.39	.01	2.93	7.64	.04	
Grade	1	.62	1.49	.01	.53	1.22	.01	.26	.69	.00	
Current use of PA	1	.01	.03	.00	.00	.00	.00	.19	.49	.00	
Agreeableness (A)	1	21.41	51.84***	.21	34.24	79.18***	.29	27.19	70.95***	.27	
Self-Rating Deviation (B)	2	10.73	25.99***	.21	12.44	28.77***	.23	5.68	14.83***	.13	
Collaboration (C)	1	15.29	37.02***	.16	13.10	30.30***	.14	9.31	24.28***	.11	
A * B	1	.32	.78	.00	.48	1.12	.01	.58	1.51	.01	
A * C	1	.10	.25	.00	.15	.34	.00	.96	2.49	.01	
B * C	1	.35	.85	.00	.03	.07	.00	.03	.09	.00	
A * B * C	1	2.4	5.81*	.03	1.82	4.20*	.02	.03	.07	.00	
Error	193	.41			.43			.38			
Total	207										
			Adjusted R ² = .46			Adjusted R ² = .47			Adjusted R ² = .37		

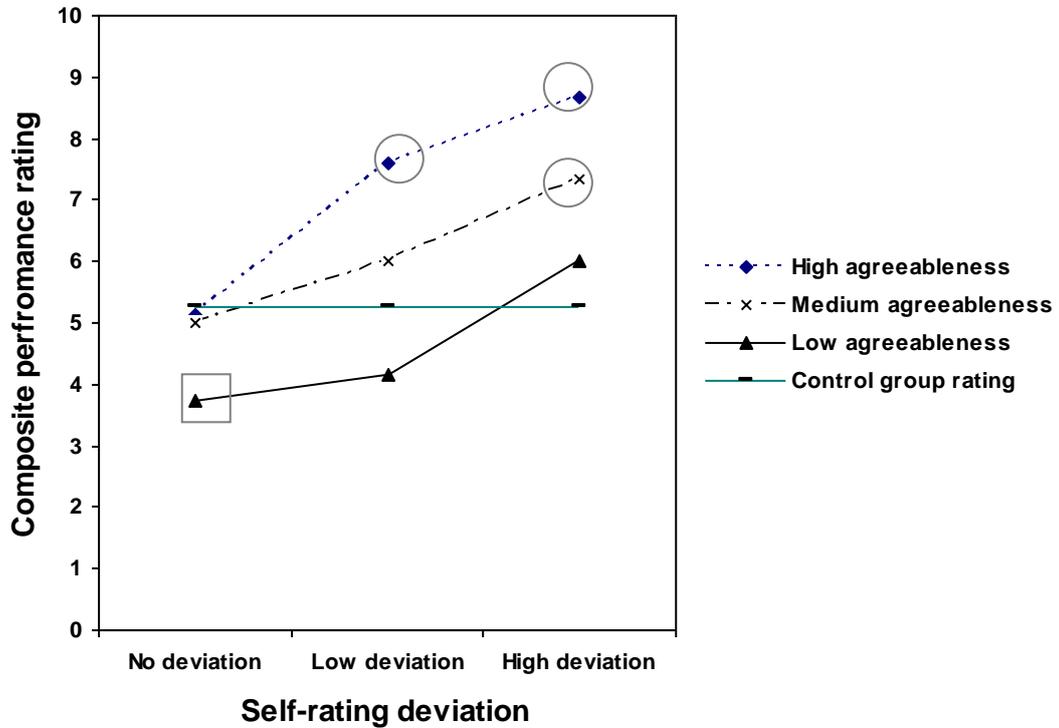
N = 207. * $p < .05$; ** $p < .01$; *** $p < .001$.

Figure Captions

Figure 1. Mean composite performance ratings given by participants with different levels of agreeableness in conditions with no prospect of future collaboration.

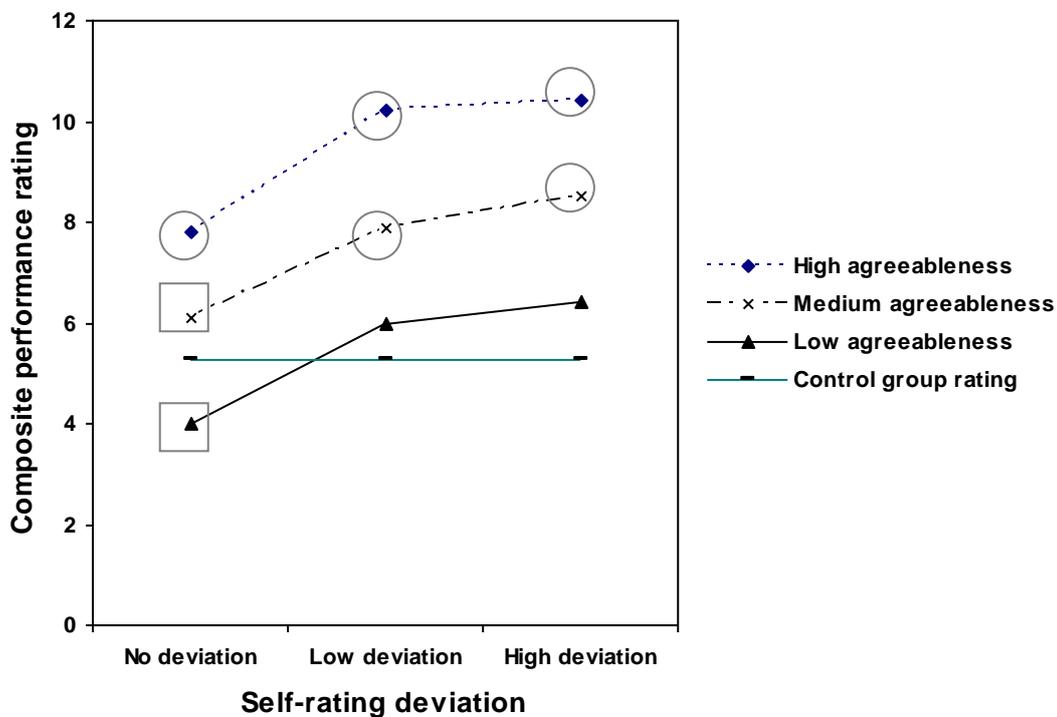
Figure 2. Mean composite performance ratings given by participants with different levels of agreeableness in conditions where there is the prospect of future collaboration.

Figure 1.



Note. Points highlighted with a □ were significantly different compared to the control group at $p < .01$; points highlighted with a O were significantly different compared to the control group at $p < .002$. This was the level of significance calculated using the Bonferroni correction to control for Type I error in the use of multiple (18) independent sample t-tests.

Figure 2.



Note. Points highlighted with a □ were significantly different compared to the control group at $p < .01$; points highlighted with a O were significantly different compared to the control group at $p < .002$. This was the level of significance calculated using the Bonferroni correction to control for Type I error in the use of multiple (18) independent sample t-tests.

Note for Editor – the note for this figure is the same as that for Figure 1. It only need be shown once if both figures are presented on the same page.

