Personnel Assessment and Decisions

Volume 1 | Issue 1 Article 5

2015

Identifying the Strongest or the Weakest Link: Effects on Subsequent Ratings

William S. Weyhrauch
Kansas State University

Satoris S. Culbertson
Kansas State University

Follow this and additional works at: https://scholarworks.bgsu.edu/pad

 Part of the Human Resources Management Commons, Industrial and Organizational Psychology Commons, and the Other Psychology Commons

Recommended Citation

Weyhrauch, William S. and Culbertson, Satoris S. (2015) "Identifying the Strongest or the Weakest Link: Effects on Subsequent Ratings," *Personnel Assessment and Decisions*: Vol. 1 : Iss. 1 , Article 5.
DOI: https://doi.org/10.25035/pad.2015.005
Available at: https://scholarworks.bgsu.edu/pad/vol1/iss1/5

This Main Article is brought to you for free and open access by the Journals at ScholarWorks@BGSU. It has been accepted for inclusion in Personnel Assessment and Decisions by an authorized editor of ScholarWorks@BGSU.

IDENTIFYING THE STRONGEST OR THE WEAKEST LINK: EFFECTS ON SUBSEQUENT RATINGS

William S. Weyhrauch[1] & Satoris S. Culbertson[1]

1. Kansas State University

ABSTRACT

This study investigated the effect of a negative designation performance rating purpose in contrast to a positive designation purpose or a deservedness purpose on (a) the ability of raters to differentiate among ratees at a later time and (b) raters' tendencies to provide subsequently more severe or lenient ratings. Results from a laboratory study involving 102 participants indicated that positive designations tend to result in subsequently lenient ratings, whereas negative designations result in severe ratings. However, the nature of a rater's previous decision had no discernable effect on the ability to differentiate levels of performance. Implications of these findings for the structuring of performance ratings procedures in contexts of short-term performance observation, such as assessment centers, are discussed.

KEYWORDS

assessment centers, performance rating, assimilation effect, rating leniency

Despite the push for systemic approaches to performance management, there remain many instances in which short-term observations of performance are rated on simple global scales, with important individual and organizational consequences. Performance ratings based on short windows of observation in which raters may or may not be familiar with a ratee's standard level of performance occur in assessment center exercises, training programs, temporary work groups, and circumstances in which employees are rated outside the regular performance management system for a one-time outcome, such as for an award or budget-driven termination.

In some cases, performance ratings represent deservedness ratings in which employees are evaluated individually regarding their worthiness of a particular outcome. Conversely, ratings may be framed in terms of designation, in which raters must identify a single candidate for an outcome (Williams, DeNisi, Meglino, & Cafferty, 1986)[1]. Given the importance of administrative outcomes, it is worthwhile to understand how specific administrative rating purposes (i.e., deservedness vs. designation) impact ratings.

Contextual factors in performance rating, such as the stated purpose of the ratings, can influence a rater's information processing/storage (e.g., Jawahar & Williams, 1997). The nature of this processing and storage may influence ratings made subsequently for a different purpose (Sumer & Knight, 1996). For example, research examining contrast and context effects within performance appraisals (e.g., Palmer, Maurer, & Feldman, 2002) suggests that a supervisor pinpointing one worker to be fired (or promoted) may lead to negative (or positive) associations and result in severe (or lenient) subsequent performance reviews. Accordingly, we sought to investigate the effect of designation versus deservedness rating purposes and outcome valence (i.e., positive or negative) on subsequent performance ratings. By identifying the ways in which ratings are unduly influenced, steps can be taken to minimize error and enhance rating quality.

We build on prior research revealing a possible assimilation effect when performance ratings have a designation purpose. Assimilation effects arise when initial ratings lead subsequent ratings to be biased towards the initial ratings (Sumer & Knight, 1996). Williams et al. (1986) found that when raters viewed performance in order to immediately designate one worker for a positive outcome, versus rating all workers on deservedness, they rated all workers higher in subsequent performance ratings. Although Williams et al. examined positive outcome designations, they did not examine *negative* designations. Instead, they simply speculated that a negative designation purpose would result in rating severity instead of leniency, indicating an assimilation effect. However, their results do not rule out the universal

1 The distinction between deservedness and designation is synonymous with the distinction between judgment and choice, respectively, in other literature. To be consistent with Williams et al. (1986), we maintain their terminology.

Corresponding author:
Dr. Satoris S. Culbertson
8A Calvin Hall, Department of Management
Kansas State University
Email: satoris@ksu.edu
Office: 785-532-6975

leniency explanation, that is, that raters tend to be more lenient when conducting administrative ratings (Greguras, Robie, Schleicher, & Goff, 2003).

This study explores the effect of a negative designation rating purpose in contrast to a positive designation purpose or a deservedness purpose on (a) the ability of raters to differentiate among ratees and (b) raters' tendencies to rate more severely or leniently. By examining a negative designation purpose, we investigate whether inflated administrative ratings can be attributed to an assimilation effect or a universal leniency effect. In practice, this question bears implications for the practical design of performance evaluation procedures, particularly in the context of a performance-contingent outcome. Better awareness of the situational factors that create severe or lenient rating bias will enable practitioners to gather the most accurate performance ratings possible, a criterion of importance to both raters and ratees.

Appraisal Purpose and Encoding

The way information is structured when it is first observed (encoding) largely determines how it is stored and later retrieved from memory (Day & Sulsky, 1995). Tulving's (1983) encoding specificity principle proposes that memory is best when retrieval conditions match encoding conditions. Similarly, the levels-of-processing framework (Craik & Lockhart, 1972) suggests that more meaningful information will be remembered more clearly due to deeper processing. A shared assertion of these theories is that encoding context affects retrieval quality.

According to the encoding specificity principle, raters with a deservedness purpose should have greater memory for performance, as the context for their observation more closely matches traditional performance rating wherein each employee is rated on the same scale. The levels-of-processing framework also suggests that deservedness raters should have greater memory because making deservedness ratings for each employee requires deeper processing than a single designation. In this case, greater memory is expected to lead to greater differentiation of worker performance levels, meaning raters with greater memory for performance will provide significantly different ratings for workers at different levels of overall task proficiency. This expectation aligns with Williams et al.'s (1986) finding that raters with a deservedness purpose were better than raters with a designation purpose at differentiating the levels of worker performance in their ratings.

Further justification for this line of reasoning comes from research on the distinction bias (Hsee & Zhang, 2004), which suggests that a joint evaluation, which involves a direct comparison of alternatives (as with a deservedness purpose), results in greater distinction between options than a separate evaluation of alternatives (as with a designation purpose). Thus, in line with past research and theory, we offer the following hypothesis.

Hypothesis 1: *Raters making designation decisions will be less able to differentiate worker performance than those making deservedness decisions.*

Effects of Previous Decisions

Previous research has established that prior decisions/judgments can exert considerable influence on subsequent ones (e.g., Sumer & Knight, 1996; Thorsteinson, Breier, Atwell, Hamilton, & Privette, 2008). The influence of previous decisions is of particular interest to organizational researchers given that repeated, but ideally independent, judgments are routine and accompanied by considerably high stakes.

In this study, we investigate whether the leniency effect demonstrated by Williams et al. (1986) replicates and, secondly, if it can be explained by an assimilation effect or the universal leniency effect resulting from shallow information encoding. Assimilation refers to rating error in the direction of an established anchor (Murphy, Balzer, Lockhart, & Eisenman, 1985; Sumer & Knight, 1996). These effects are also referred to as context effects (Kravitz & Balzer, 1992; Palmer et al., 2002), referring to the influence of the context (anchor) on the distribution of ratings, independent of what is being rated. An assimilation effect would suggest that an initial positive or negative decision would result in lenient and severe subsequent ratings, respectively. One theoretical explanation of the assimilation effect is the priming hypothesis (Collins & Quillian, 1969), which purports that cognitive categories (e.g., effective performance) used to organize the perception of one worker will prime the use of these categories in the perception of subsequent workers. In essence, thinking of an initial worker's effective performance will produce benefits for subsequent workers by priming the rater to think positively. This is essentially what Shafir, Simonson, and Tversky (1993) are referring to in subsection 2 "Reasons Pro and Con" (pp. 15-18) of their paper on reason-based choice, that the results of a binary choice are influenced by the framing of the choice as endorsement of one versus rejection of one.

Williams et al. (1986) found what appears to be an assimilation effect that might be explained by the priming hypothesis. Specifically, they found that raters given a positive designation purpose subsequently gave more lenient ratings than raters given a deservedness rating purpose. One explanation for this is that the designation purpose limits the amount of performance information retained in memory for each worker because it does not require as much processing as the deservedness purpose. The designation purpose requires less processing because there is no need to differentiate all levels of proficiency, just the best from the rest; whereas, the deservedness purpose forces raters to evaluate each worker's individual performance.

Less processing, however, leads to a lack of memory for performance that may have a variety of effects on ratings, depending on the cause. If the leniency noted by Williams et al. (1986) was caused by assimilation, then a negative designation purpose would lead to severity. Conversely, it may be that lack of memory for performance results in leniency. Leniency may result from both positive *and* negative designation decisions due to shallow processing. Designation requires limited cognitive processing because there is no need to differentiate all levels of proficiency, merely the most extreme. Shallow processing, however, inhibits memory for performance. Insufficient memory for performance may then result in leniency, regardless of whether an outcome is positive or negative. Researchers have shown that administrative ratings tend to be more lenient (Greguras et al., 2003; Jawahar & Williams, 1997) than developmental ratings. This universal leniency effect may be exacerbated for designation ratings due to the lower processing requirements.

The assimilation and universal leniency explanations result in the same prediction for a positive designation purpose (i.e., significantly higher ratings than a deservedness purpose) but different predictions for a negative designation purpose. Although there is probably a stronger theoretical case to be made for the assimilation hypothesis, this cognitive phenomenon may not be strong enough to overcome the tendency toward leniency in administrating rating situations. As such, we offer the following competing hypotheses regarding these two perspectives.

Hypothesis 2a: *Raters with a positive designation purpose will give significantly higher ratings than those with a deservedness purpose or negative designation purpose.*

Hypothesis 2b: Raters with a positive or negative designation purpose will give significantly higher ratings than those with a deservedness purpose.

METHOD

Participants

Participants were 102 undergraduate students (56 women, 46 men) from a Midwestern university participating for course credit. The sample was ethnically diverse (57% Caucasian, 13% African-American/Black, 11% Asian, 9% Hispanic), with a mean age of 20.3 (SD = 3.7). A large majority were either employed part time (n = 48) or had employment experience (n = 45).

Procedure and Materials

Participants were randomly assigned across four conditions: negative designation, positive designation, deservedness, or a control condition. Participants were (deceptively) informed they were providing evaluations for selecting CPR-capable participants for a research study with significant monetary compensation. All participants viewed an 8-minute CPR instruction video featuring a demonstration and explanation of the proper technique and steps in administering CPR to an adult and an infant. They were also provided with written guidelines of what is and is not correct CPR. Though there are many important steps to take in an emergency situation (e.g., calling 911), chest compressions and breaths are the crucial elements in keeping a person alive. As such, these four tasks (adult chest compression, adult breaths, infant chest compression, and infant breaths) were chosen as the most appropriate way to divide CPR into distinct tasks. All participants then viewed videos of four similar female confederates performing the tasks: a high performer (75% proficiency, i.e., 3 of 4 tasks performed correctly), two medium performers (50% proficiency), and one poor performer (25% proficiency). Performance failure on a task was operationalized as an applicant making a clear error on a task but not necessarily doing everything wrong on that task. To avoid order effects, four versions of the performance video were made, such that each applicant was shown in each position (i.e., first, second, third, or fourth) once. The four versions of the video were randomized across sessions. Also, each applicant was female, in her early-mid twenties, and of a similar body type and attractiveness to avoid biases associated with applicant gender, age, or physical appearance. An additional issue in developing the performance videos was whether mistakes on infant tasks might be perceived as more serious than mistakes on adult tasks. Thus, the most proficient applicant made her mistake doing adult compressions. One middle proficiency applicant erred on the adult compressions and infant breaths. To balance across this proficiency level, the other middle proficiency applicant made mistakes on adult breaths and infant compressions. Finally, the lowest proficiency applicant made a mistake on all tasks except for the infant compressions.

Participants were given varying instructions for how to observe performance, consistent with their condition. Those in the deservedness condition were instructed to observe performance in order to rate all participants on a 7-point scale of how much they deserve to be selected. Those in the designation conditions were instructed to observe in order to identify a single best or worst performer. Those in the control condition had no instructions and rated all participants afterward with the 7-point scale.

Similar to Williams et al. (1986), all participants returned 2 days later to the same room for Session 2 of the study. Some limited attrition (~10%) occurred between sessions but was not systematically related to any of the conditions. Relying on their memory and a photo of each confederate, all Time 2 participants made overall performance ratings for each confederate on a 7-point scale from 1 = *poor* to 7 = *outstanding*. Each level of proficiency (25%, 50%, 75%) was plotted on this 7-point rating scale; specifi-

TABLE 1.
Mean Performance Ratings for Each Proficiency Level as a Function of Appraisal Purpose

Applicant proficiency (% correct)	Deservedness	Positive designation	Negative designation	Control	Total
75%	5.07 (1.14)	5.89 (.97)	4.63 (.74)	4.43 (.94)	5.04 (1.27)
50%	3.93 (1.13)	4.52 (.85)	3.35 (1.02)	3.81 (.96)	3.91 (1.07)
25%	3.19 (1.24)	3.48 (1.37)	2.07 (1.21)	3.19 (1.08)	2.97 (1.34)
Total	4.03 (.85)	4.60 (.74)	3.35 (.80)	3.81 (.75)	3.96 (.90)

Notes. $N = 102$. (27 in each experimental condition, 21 in control condition); Mean overall performance ratings (scale of 1 to 7) with standard deviations in parentheses. At the 75% proficiency level, significant differences exist only between the positive designation condition and all other conditions. At the 50% proficiency level, the only significant difference is between the positive and negative designation conditions. At the 25% proficiency level, significant mean differences exist between the positive and negative designation conditions, the negative designation and deservedness conditions, as well as the negative designation and control conditions. True scores were set at the quartile points of the 7-point rating scale, 5.25 (75% proficiency), 3.5 (50% proficiency), and 1.75 (25% proficiency). These true scores provide some indication of which mean ratings may be lenient/severe in reference to an absolute (as opposed to relative) standard.

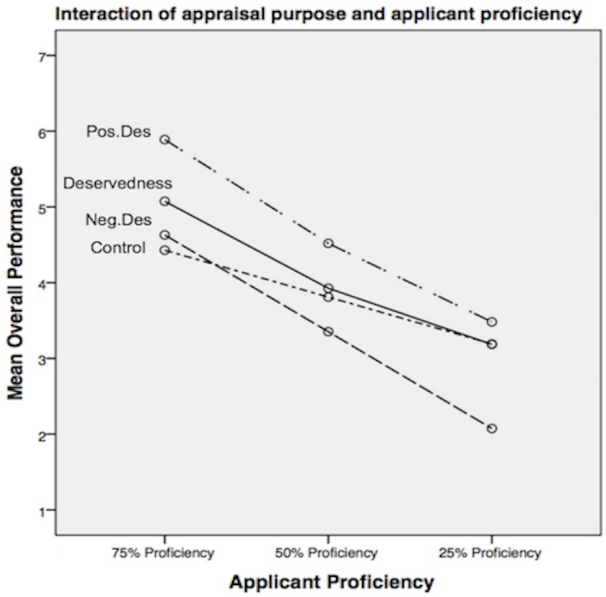

FIGURE 1. Appraisal purpose and applicant proficiency. Within appraisal purpose condition, ratings for each proficiency level were significantly different. In the deservedness condition, the 25% and 50% proficiency levels were not significantly different. In the control condition, only the 75% and 25% proficiency levels were significantly different.

cally, the 25% proficiency level translates to a score of 1.75 (25% of the maximum score) on the performance scale, 50% translates to 3.5, and 75% translates to 5.25.

RESULTS

Session 2 mean overall performance ratings are presented in Table 1. A 4 (purpose) × 3 (proficiency) mixed factorial ANOVA was conducted, treating proficiency as a within-subjects variable and appraisal purpose as a between-subjects variable. Post-hoc comparisons were calculated with Tukey's HSD. The main effect of proficiency, $F(2,208) = 224.082$, $p < .001$, $\omega^2 = 0.58$, and post-hoc comparisons demonstrated that raters perceived different levels of performance, as designed. Specific comparison statistics are available from the authors.

Hypothesis 1 predicted that deservedness raters would better differentiate levels of proficiency than would designation raters. This was tested by the interaction of purpose and proficiency, $F(6,208) = 3.662$, $p < .05$, $\omega^2 = 0.01$, which indicated that rating purpose influenced participants' ability to differentiate proficiency levels (see Figure 1). However, simple effects analysis revealed that raters in all conditions significantly differentiated each level of performance. Therefore, Hypothesis 1 was not supported, thereby failing to replicate Williams et al. (1986).

Hypotheses 2a and 2b examined the pattern of ratings among positive and negative designation raters as well as

deservedness raters, predicting higher ratings only for positive designation (Hypothesis 2a) or among both designation conditions (Hypothesis 2b). The main effect of purpose, $F(3,104) = 14.344$, $p < .001$, $\omega^2 = 0.11$, was probed by examining Tukey's HSD, which revealed significantly higher ratings in both the positive designation and deservedness conditions. This evidence is contrary to Hypothesis 2b, which is based on the universal leniency hypothesis, and provides further support to the assimilation hypothesis underlying Hypothesis 2a. A simple effects analysis across proficiency levels demonstrated the consistency of this effect acrossproficiency levels. Positive designation ratings were lenient relative to true scores for each proficiency level. Conversely, negative designation ratings were severe relative to true scores in the top two proficiency levels.

DISCUSSION

This study investigated an important unresolved question in the performance rating literature regarding the influence of a designation purpose on performance ratings. Specifically, this study examined two competing predictions for the effect on subsequent performance ratings when performance is observed under the frame of a positive or negative designation decision. This analysis provides additional context for the findings of Williams et al. (1986) and identifies a boundary condition for when administrative ratings will tend to be lenient. Our results provide new insights on an important phenomenon in performance rating: the influence of the decision context. Specifically, we found that negative designations have a different effect on ratings than do positive designations, thereby indicating assimilation as the cause of leniency resulting from positive designations. As hypothesized, the negative designation purpose resulted in severe ratings, relative to the deservedness purpose. This not only contributes new understanding to the literature on the performance rating cognitions but also further clarifies the underlying process behind the findings of Williams et al. (1986), namely that assimilation drives the effect, rather than leniency due to shallow processing.

We were also able to test the generalizability of the effects found by Williams et al. (1986) with a different task. Consistent with Williams et al., raters observing performance to make a positive designation gave higher subsequent ratings than those given a deservedness purpose. However, our findings failed to replicate their finding that raters with a designation purpose failed to differentiate all levels of proficiency. On the contrary, raters in all conditions were able to differentiate each proficiency level from the others. Finally, consistent with Williams et al., lenient ratings resulted when raters were given a positive designation purpose compared to when they were given a deservedness rating purpose.

Although our findings did not support a difference in recall between designation and deservedness conditions, it may be that when raters are more familiar with the targets and have preexisting notions of their performance, their ability to distinguish performance levels within a particular performance sample is diminished. However, we did expect to replicate Williams et al.'s (1986) findings. This may result from errors in our performance videos being relatively more apparent. Furthermore, the deception employed and the serious nature of the task (CPR vs. woodworking) may have led our participants to take their task more seriously and thus retain more performance information. Further research employing an examination of the actual recall of performance details at Time 2 would provide further clarity on the results of Hypothesis 1 and whether designation purposes result in limited memory for performance due to shallow processing.

A final interesting observation of our data can be seen when comparing the range of scores for the designation conditions versus the deservedness conditions. As shown in Table 1, the range between the top and bottom performer is substantially larger for the designation conditions (5.89 - 3.48 = 2.41 and 4.63 - 2.07 = 2.56) compared to the deservedness condition (5.07 - 3.19 = 1.88). Although it is unclear the reason for this difference, it could be that raters making designation decisions were more impacted by post-decisional dissonance than were raters making deservedness ratings and as such felt more compelled to reduce their dissonance by increasing the difference between the best and worst candidates. More research is needed to explore this possibility, as well as the ways in which such dissonance, if it exists, can be reduced without biasing ratings[2] .

Strengths and Limitations

This study employed a strong experimental design focused on two competing explanations for a potentially very high stakes phenomenon. This phenomenon has wide practical relevance to performance rating contexts, including assessment centers, performance appraisals, and skill-based certification tests. Many steps were taken to solidify internal validity and eliminate confounds. Nevertheless, our study is limited by the laboratory setting and generalizability of CPR to more traditional work tasks, as well as the potential for a stronger leniency effect in field samples. Furthermore, our study does not address the effect of multidimensional performance ratings. Last, our estimate of true scores relies on the potentially faulty assumption that deservedness decisions are a linear function based on absolute standards. It is possible that deservedness is nonlinear, however, with individuals making ratings based more on comparative standards rather than based on an absolute standard. Future researchers should explore these issues.

2 We wish to thank an anonymous reviewer for making this observation.

Implications

Practitioners in applied settings of performance evaluation should take careful note of these findings. This study has important implications for the design of performance rating procedures in one-time performance observation and assessment scenarios in which targets are not well-known to the rater, such as assessment centers and skill-based certification programs. These contexts often have particularly high stakes for the ratee, more so than a routine performance management review. Our findings suggest that, even if performance ratings are primarily used for designation purposes (e.g., identifying those with the highest potential), each employee should be evaluated equally on their deservedness for that outcome. Raters should carefully balance both positive and negative aspects of performance rather than exclusively focusing on how far a target is from being ideal or problematic.

These findings should be of interest to a broad audience of academics and practitioners alike who up to this point have accepted the notion that all ratings for administrative purposes are lenient. On the contrary, our results suggest that when ratings are conducted for the purposes of a negative outcome, rating error shifts toward severity. Researchers and consultants who rely on research regarding performance rating purpose effects should be aware of this new evidence.

CONCLUSION

Our study presents strong, experimentally derived evidence to a practically relevant question and challenges a commonly held belief among scholars regarding the effect of administrative performance ratings. We demonstrated that raters who are evaluating with a positive or negative designation in mind are likely to give lenient or severe ratings overall, respectively. As such, we have confirmed the importance of structuring performance rating procedures to maximize the amount and quality of performance information retained by raters.

REFERENCES

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-248.

Craik, F. I. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. Journal of Verbal Learning & Verbal Behavior, 11, 671-684.

Day, D. V. & Sulsky, L. M. (1995). Effects of frame-of-reference training and information configuration on memory organization and rating accuracy. *Journal of Applied Psychology, 80*, 158-167.

Greguras, G. J., Robie, C., Schleicher, D. J., & Goff III, M. (2003). A field study of the effects of rating purpose on the quality of multisource ratings. *Personnel Psychology, 56*, 1-21.

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology, 86*, 680-695.

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

Kravitz, D. A., & Balzer, W. K. (1992). Context effects in performance appraisal: A methodological critique and empirical study. *Journal of Applied Psychology, 77*, 24-31.

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70*, 72-84.

Palmer, J. K., Maurer, T. J., & Feldman, J. M. (2002). Context and prior impression effects on attention, judgment standards, and ratings: Contrast effects revisited. *Journal of Applied Social Psychology, 32*, 2575-2597.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49*(1), 11-36.

Sumer, H. C., & Knight, P. A. (1996). Assimilation and contrast effects in performance ratings: Effects of rating the previous performance on rating subsequent performance. *Journal of Applied Psychology, 81*, 436-442.

Thorsteinson, T. J., Breier, J., Atwell, A., Hamilton, C., & Privette, M. (2008). Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes, 107*, 29-40.

Tulving, E. (1983). *Elements of episodic memory*. Oxford, UK: Oxford University Press.

Williams, K. J., DeNisi, A. S., Meglino, B. M., & Cafferty, T. P. (1986). Initial decisions and subsequent performance ratings. *Journal of Applied Psychology, 71*, 189-195.

RECEIVED 1/16/15 **ACCEPTED** 8/24/15