Between-individual comparisons in performance evaluation: A perspective from prospect theory

Kin Fai Ellick Wong
Hong Kong University of Science and Technology

Jessica Y. Y. Kwong
Chinese University of Hong Kong

Author’s Notes

We thank Virginia Unkefer and two anonymous reviewers for their helpful comments on an early version of this article. We also thank Kiniu Wong for inspiring us to do this research. We are grateful to Edith Li, Phil Leung, and Peggy Chan for their assistance in participant recruitment and data analyses. Correspondence concerning this article should be addressed to Kin Fai Ellick Wong, Department of Management of Organizations, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR. E-mails may be sent to mnewong@ust.hk
Abstract

This paper examines how between-individual comparisons influence performance evaluations in rating tasks. We demonstrated a systematic change in the perceived difference across ratees as a result of changing the way performance information is expressed. Study 1 found that perceived performance difference between two individuals was greater when their objective performance levels were presented with small numbers (e.g., absence rates of 2% vs. 5%) than when they were presented with large numbers (e.g., attendance rates of 98% vs. 95%). Extending this finding to situations involving trade-offs between multiple performance attributes across ratees, Study 2 showed that the relative preference for one ratee over another actually reversed when the presentation format of the performance information changed. We draw upon prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981) to offer a theoretical framework describing the between-individual comparison aspect of performance evaluation.
Between-individual comparisons in performance evaluation: A perspective from prospect theory

One day, one of the authors overheard two students discussing the free throw performance of some NBA players. They tried to classify the players by giving them a grade from A+ to E. They mentioned Reggie Miller and agreed that he was an excellent free throw shooter and hence should be given an A grade. Then, they discussed Mike Bibby and had a disagreement. One said, “Mike’s career free throw percentage is 80% and Reggie’s percentage is 89%. Their performances do not differ much. I think Mike and Reggie should be classified with the same grade, though Reggie is better. So, I give Mike an A-.“ However, the other said, “No, Reggie missed only 11 free throws per 100. The miss percentage by Mike is 20%, which is almost double Reggie’s. How can they be classified with the same grade? I think a B+ or even a B is more appropriate for Mike.”

The format of performance evaluation can be classified either as a relative rating system or an absolute rating system (see Cascio, 1998). Relative rating systems ask raters to compare an employee’s performance with the performance of other employees. Examples include simple ranking, paired comparisons, and forced distribution. Absolute rating systems ask raters to judge employees’ performance on the basis of comparing their performance with performance standards, independent of between-individual comparisons. Examples include critical incidents, behavior checklists, and graphic rating scales.

While making judgments about performance under absolute rating systems is supposed to be mainly determined by the discrepancy between the observed or recalled behavior and an absolute performance standard defined in subjective or objective terms (i.e., a behavior-standard
discrepancy) (Feldman, 1986; Murphy & Cleveland, 1995), a rating score may also reflect between-individual comparisons. As illustrated by the above scenario, the performance rating of one player may be influenced by how he/she compares with others. Research on between-individual contrast effects has shown that evaluation of the current ratee is often contrasted away from the evaluation of the preceding ratees (Maurer & Alexander, 1991; Wexley, Sanders, & Yukl, 1973; Wexley, Yukl, Kovacs, & Sanders, 1972). In organizations, the influence of between-individual comparisons in evaluations appears to be salient because (a) a typical evaluation often involves multiple ratees (Bernardin & Villanova, 1986), and (b) between-individual comparison is a common purpose of performance appraisal (e.g., promotion, bonus allocation, and salary change, etc.; Cleveland, Murphy, & Williams, 1989). This paper takes a perspective from the judgment and decision-making literature to understand the mechanism through which between-individual comparisons influence performance evaluation.

Prior conceptual development of performance evaluation has primarily focused on the behavior-standard discrepancy. The cognitive approach to performance appraisal (DeNisi, Cafferty, & Meglino, 1984; Landy & Farr, 1980) stresses performance evaluation as reflecting the “true scores” of performance (Wherry & Bartlett, 1982). Feldman (1986) introduces the concept of the “performance model” to describe the standards used in evaluating performance. In their organizational, goal-based approach to performance appraisal, Murphy and Cleveland (1995) characterize performance judgment as a result of the comparison process that detects the behavior-standard discrepancy. Some other models also proposed that performance judgments are made with reference to implicit or explicit standards (e.g., DeCottis & Petit, 1978; Ilgen, 1983).

Despite the many contributions these models have made to understanding how raters judge and rate individual performance, their fundamental focus is on behavior-standard comparisons.
Little is known about how performance evaluation may be influenced by between-individual comparisons. In this paper, we address this issue by borrowing concepts from prospect theory in judgment and decision-making research (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981). We are particularly interested in the mechanisms through which raters perceive performance differences across individuals. We argue that the same information when presented in different ways may amplify or diminish the perceived performance differences. On the theoretical side, this research may place another stone on the path toward a comprehensive view of performance evaluation. On the practical side, this paper may promote people’s awareness that the way they present performance data may have a significant impact on how two candidates are evaluated.

Prospect Theory and Hypotheses

We first give a brief overview of prospect theory, followed by our hypotheses on perceived performance differences across individuals derived from prospect theory. We then report on two studies designed to examine these hypotheses.

Prospect theory

Prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981) describes the decision processes involved when people make comparisons between options. There are two important components in this theory, frame of reference and subjective value function. Figure 1 presents a typical subjective value function with a neutral reference point that prospect theory postulates.

Insert Figure 1 about here
Frame of reference. Prospect theory posits that the selection of a reference point is crucial to option evaluation. A reference point is the point (an internal standard) with which people compare the objective value of an option so as to classify the option as something positive (i.e., when the option’s value is larger than the reference point) or as something negative (i.e., when the option’s value is smaller than the reference point). Objectively identical options can be framed positively or negatively, such that an option perceived as a gain in one frame will be perceived as a loss in another frame. A positively framed option is thought to decrease one’s evoked reference point, whereas the same information framed negatively is thought to increase the evoked reference point (Abelson & Levi, 1985; Highhouse & Paese, 1996).

For example, in a scenario describing that 600 people are infected by an unusual disease, the effectiveness of a medical program could be framed in a positive way (e.g., 200 people will be saved). This framing would lead people to presuppose loss of all lives, which would evoke a relatively low reference point (i.e., 0 people will be saved). However, framing the same program negatively (i.e., 400 people will die) would lead people to presuppose no lives lost, which would evoke a relatively high reference point (i.e., 600 people will be saved, or 0 people will die). This shift in the reference point across the two frames leads to a perception that the medical program is more attractive (i.e., it has a higher subjective value) when it is positively framed than when it is negatively framed. The program in the positive frame will lead to a sense of saving 200 lives when compared to the “loss of all lives” reference point. Conversely, the negative frame will lead to a sense of losing 400 lives when compared to the “no lives lost” reference point. A treatment that will save 200 lives is more desirable than one that will cause 400 deaths. Thus, the medical program is more attractive when it is presented in a positive frame than when it is presented in a negative frame.

Framing effects like this has been shown to be very robust. Similar findings that perceived
favorability of an option depends on framing have been demonstrated in a wide variety of social and personal settings (for a review, see Levin, Schnedier, & Gaeth, 1998), including layoffs (Brockner, Wiesenfeld, & Martin, 1995), purchase of products (Levin & Gaeth, 1988), escalation situations (Davis & Bobko, 1986), project funding allocation (Dunegan, 1993), medical treatments (Levin, Schnittjer, & Thee, 1988), and condom use (Linville, Fischer, & Fischhoff, 1993).

We acknowledge here that the term “frame-of-reference” used in prospect theory and in this paper should be distinguished from the term “frame-of-reference training” used in the literature on performance appraisal (Bernardin & Buckley, 1981; Day & Sulsky, 1995; Pulakos, 1984, 1986). Frame-of-reference training refers to a program that is designed to increase rating accuracy by “calibrate(ing) raters so that particular scores on individual rating dimensions have at least roughly equivalent meanings for all raters.” (Murphy & Cleveland, 1995, p. 201) That is, all raters are trained to map a particular performance level to a particular rating. Instead of introducing specific standards for evaluation purposes, frame-of-reference used in prospect theory refers to the introduction of a reference point that makes a frame casting “the same critical information in either a positive or negative light” (Levin et al., 1998, p. 150). Thus, whereas the term frame-of-reference in prior performance appraisal literature describes calibration of a common set of internal reference across different raters to evaluate performance, in this paper it refers to the shifts of the internal reference point as a result of the ways in which the same information is presented (positive versus negative frames).

Subjective value function. Under the framework of prospect theory, an option is associated with a subjective value expressed as positive or negative deviations (i.e., gains or losses) from a neutral reference point (i.e., with a subjective value of zero). The hypothesized value function follows the non-linear perception and judgment function in many sensory and
perceptual dimensions (Stevens, 1962), such as loudness (Stevens, 1936), brightness (Stevens, 1961), and weight perception (Engen, 1972; Gescheider, 1976). The value function is S-shaped (see Figure 1), concave in the positive domain (i.e., above the reference point) and convex in the negative domain (i.e., below the reference point). Specifically, as the gain increases in the objective value, the increase in the subjective value (or the slope in the graph) decreases. For example, the difference in the subjective value between gains of $10 and $20 appears to be greater than that between gains of $110 and $120. As the loss increases in the objective value, the drop in the subjective value (or the slope in the graph) decreases. For example, the difference in subjective value between losses of $10 and $20 appears to be greater than that between losses of $110 and $120. Consistent with the S-shaped value function, previous studies have shown that the marginal subjective values of money, non-monetary events (e.g., stepping on gum), and objects (e.g., the Irish Sweepstakes) with both gains or losses decrease as their sizes increase (Bernoulli, 1954; Galanter, 1990; Galanter & Pliner, 1974).

Despite the robustness of this S-shaped value function, literature that explains this pattern is rather sketchy. It has been proposed that it is because human beings have evolved to possess this characteristic. One explanation offered by Arkes (1991) is that the values function evolved to be non-linear because of its superior adaptive merit over a linear system. From the perspective of evolution, a physiological system that translates external stimuli onto psychological responses in a way follows non-linear functions is more economical than a linear transduction system. In explaining the S-shaped value function of prospect theory and other psychological power functions, Arkes (1991) noted,

“A system that translated physical intensity in a linear manner onto psychological responses would impose an immense cost on any transduction system. Extreme stimuli, which occur
relatively infrequently, would have to be coded with as great a level of discriminability as the more frequent middle-range stimuli. Any nonlinear system with an asymptote at the extreme end (or ends) would have the benefit of eliminating the structures and processes needed to discriminate small changes in rare events, such as very intense sounds or extremely heavy weights. Of course, sacrificing discriminability at the ends of the continuum has a cost." (p. 490).

Performance comparison and prospect theory

We propose that the concepts of frame of reference and subjective value function from prospect theory provide significant insights into the between-individual comparison in performance evaluation. Specifically, we hypothesize that perceived performance (i.e., subjective performance) is guided by the subjective value associated with a standing on the behaviors constituting the performance (i.e., objective performance). Each standing on objective performance corresponds to a subjective value. Objective performance often could be expressed positively (e.g., attendance rate) or negatively (e.g., absence rate). The relationship between objective performance and subjective value is S-shaped, concave in the positive domain and convex in the negative domain (see Figure 2). Thus, as the score in the positively framed domain increases (i.e., absolute performance gets better), the increase in the subjective value decreases, whereas as the score in the negatively framed domain increases (i.e., absolute performance gets worse), the drop in the subjective value decreases.

--------------

Insert Figure 2 about here

--------------

The idea of the frame of reference in prospect theory implies that the same behavior framed
positively or negatively may lead to different subjective values. Extrapolating from prospect theory, a rater should presuppose a higher reference point when the objective behavior is framed negatively than when it is framed positively. The same level of performance therefore is likely to be perceived as better when it is described in a positive manner than in a negative manner. For example, expressing “participation” positively with attendance rate (e.g., the attendance rate is 93%) would lead one to presuppose that a relatively low level of participation is normal (i.e., a low reference point). Expressing the same behavior negatively with absence rate (e.g., the absence rate is 7%) would lead one to presuppose that a relatively high level of participation is normal (i.e., a high reference point). Thus, it is likely that an attendance rate of 93% is perceived to be more favorable than an absence rate of 7%. Accordingly,

**Hypothesis 1.** The same level of objective performance will be perceived as better when it is positively framed than when it is negatively framed.

One implication of the subjective value function shown in Figure 2 is that the perceived performance difference between two persons should be greater when their objective performance is close to the central zero point (i.e., small numbers) than when it is far away from it (i.e., large numbers). This is because the slope is steeper when the objective performance is close to zero. Note that the same objective performance could be represented as a standing point close to or far away from the central zero point, depending on how it is expressed. For example, absence rates of 7% and 3% are identical to attendance rates of 93% and 97%, respectively. Yet, the perceived difference looms larger in the former than in the latter pair because, as shown in Figure 2, the slope between 93% and 97% is flat but that between 7% and 3% is steep. The difference in the subjective value between attendance rates of 93% and 97%, therefore, appears to be smaller than
that between absence rates of 7% and 3%. Thus, the S-shaped characteristics of the subjective value function lead to the following hypothesis:

**Hypothesis 2.** The perceived performance difference between individuals is smaller when the objective behaviors are expressed with large absolute numbers than when they are expressed with small absolute numbers.

Note that Hypothesis 2 should also hold when small numbers are associated with a positive attribute and large numbers are associated with a negative attribute. That is, the perceived performance difference between individuals suggested in Hypothesis 2 is primarily driven by the magnitude of the numbers (large versus small) rather than the positive or negative frame of the attribute (e.g., absence versus presence). We return to this point in Study 2.

The above reasoning and hypotheses may have a strong practical implication with respect to rating accuracy in terms of **differential elevation**, which refers to rating accuracy in discriminating among ratees after averaging over all performance attributes (Cronbach, 1955). For example, poor differential elevation results when the rater drastically over- or under-estimates the extent to which different ratees’ overall performance actually differs (Murphy & Cleveland, 1995). Note that this type of accuracy is particularly important in personnel practices because the majority of personnel decisions, such as promotion and salary increases, involve differential elevation (Murphy & Cleveland, 1995; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

We further hypothesize that, in extreme cases, preference reversal may occur between two ratees with a trade-off between two or more attributes that are relevant to their performance. To illustrate, consider the case when you need to select between Andy or David for a programmer position in your company (see Table 1). In this case, Andy has better programming knowledge
but has a poorer academic record. When programming knowledge is expressed with large numbers (e.g., programming success rate), the perceived difference between Andy and David in this dimension may become insignificant. According to prospect theory, the overall rating of David is likely to be higher than that of Andy because the rater would base his/her rating primarily on the difference in their academic records. However, when programming knowledge is expressed with small numbers (e.g., programming failure rate), the perceived difference between Andy and David in this dimension appears to be large. Because the rater would now consider the “substantial” difference in programming knowledge between the two ratees, the overall rating of Andy is more likely to be higher than that of David as compared with the previous case. Therefore, an attribute may become less influential when it is expressed with large numbers than when it is expressed with small numbers, leading to reversals in the overall rating of the two candidates. We therefore posit that:

**Hypothesis 3:** Preference reversal will occur in the overall evaluation of two ratees with a trade-off between two performance attributes when the performance is displayed differently. By switching the expression of the two performance attributes (large versus small numbers), one ratee may be more preferred than the other ratee in one condition yet less preferred in the other condition.

-------------

Insert Table 1 about here

-------------

**Overview**

We tested the above hypotheses with two studies. Participants in Study 1 (undergraduate and graduate students) were asked to act as managers who were to conduct performance
evaluations and those in Study 2 (professional teachers) were asked to conduct performance
evaluations to aid school principals in making a hiring decision. They were presented with
various performance attributes (e.g., absence rate, failure rate, etc.) of different ratees. Study 1
examined if perceptions of performance change when the performance is framed positively
versus negatively (i.e., Hypothesis 1) and if the perceived performance difference between two
ratees depended on the numbers (large or small) used to express their performance (i.e.,
Hypothesis 2). Study 2 examined the possible preference reversal in rating (i.e., Hypothesis 3)
by presenting information with trade-offs between two performance attributes of two ratees.

Study 1

Method

Participants, design, and evaluation task. Eighty undergraduate and graduate students
participated in this study. Three cases were excluded from analyses because of incomplete
information. This experiment used a 2 x 2 x 6 mixed design, with Performance Framing
(positive vs. negative) as a between-subjects factor, Ratee (Andy vs. William) and Rating
Dimensions (attendance, punctuality, accuracy, completeness, success, and overall performance)
as within-subject factors. Respondents were randomly given one of the two scenarios in which
the performance-framing manipulation was embedded. Participants were told that they were “a
sales manager of a telecom company” and were recently required to evaluate the performance of
three subordinates (Andy, David, and William) on the basis of five performance dimensions.
Andy performed better than William on all attributes. David was presented as a filler case (i.e., a
dummy case in which the information on David was irrelevant to our manipulation) to increase
the perceived variability across the ratees. Ratings of David were not used in the analyses.

Details of the performance attributes by the two framing conditions are summarized in Table
2. In the negative framing condition, the attributes were expressed in a negative direction, associated with small numbers (i.e., < 25%). In the positive framing condition, the attributes were expressed in a positive direction, associated with large numbers (i.e., > 75%). The cutoffs (i.e., < 25% for small numbers) were to some extent arbitrary, but we attempted to control them within a realistic range for an employee. More specifically, the attendance attribute was expressed in terms of the absence rate (percentage of being absent in the past year) versus the presence rate (percentage of being present in the past year) in the two conditions. Punctuality was expressed in terms of the rate of tardiness (percentage of being late in the past year) versus the rate of being on time (percentage of being on time in the past year). Accuracy was expressed in terms of the error rate (percentage of making errors in invoices in the past year) versus the accuracy rate (percentage of not making errors in invoice in the past year). Completeness was expressed in terms of the incompleteness rate (percentage of not completing the work assigned by supervisors) versus completeness rate (percentage of completing the work assigned by supervisors). Success was expressed in the customer-termination rate (percentage of customers terminating the telecom service after signing an initial agreement) versus the customer retention rate (percentage of customers joining the telecom service after signing an initial agreement).

Note that regardless of the framing of the attribute, the performance data always referred to same objective facts; for example, an absence rate of 5% was objectively identical to a presence rate of 95%.

Insert Table 2 about here

Procedure. Participants were asked to evaluate the subordinates by indicating how satisfied they were with each subordinate on six rating dimensions: satisfaction with each of the
five attributes and an overall satisfaction with the ratee. Ten-point Likert scales were used, with 1 = extremely dissatisfied and 10 = extremely satisfied. We used a 10-point scale here because of its popularity in performance evaluation (Kaplan & Saccuzzo, 2001). We were aware that it just exceeded the optimal scale points for Likert scales (Bass, Cascio, & O’Connor, 1974). This issue will be addressed by using a 7-point scale in Study 2.

All participants rated Andy first, then David, and finally William. All performance information was available to participants throughout the experiment.

Results

Figure 3 summarizes the results of this study. Two obvious patterns are observed. First, this figure clearly shows that across all six rating dimensions, the mean ratings in the positively framed condition are higher than those in the negatively framed condition. Second, the rating differences between Andy and William appear to be greater in the negatively framed condition than in the positively framed condition.

The above observations were confirmed by a 2 (Framing) by 2 (Ratee) by 6 (Rating dimensions) analysis of variance (ANOVA). The main effect of Framing was significant, $F (1, 75) = 43.7, \text{MSE} = 13.33, p < .0001, \eta^2 = .37$, indicating that the mean rating in the positively framed condition (7.33) was greater than that in the negatively framed condition (5.74). This finding supports Hypothesis 1 that the same objective performance is perceived as better when it is framed positively than when it is framed negatively. The main effect of Ratee was also significant, $F (1, 75) = 190.24, \text{MSE} = 3.44, p < .0001, \eta^2 = .72$, indicating that the mean rating for Andy (7.38) was greater than that for William (5.70). The main effect of Rating dimensions
APA IN PRESS ARTICLE: If you plan to quote this manuscript, be sure to check the actual wording of the final published article.

was also significant, $F(5, 375) = 56.54$, $MSE = 1.98$, $p < .0001$, $\eta^2 = .43$, indicating that there were differences in the ratings across the dimensions.

More important, the Framing × Ratee interaction was significant, $F(1, 75) = 10.55$, $MSE = 3.44$, $p < .005$, $\eta^2 = .12$, indicating that the mean rating differences between Andy and William were smaller in the positively framed condition (7.97 vs. 6.69) than in the negatively framed condition (6.78 vs. 4.70). Because the positively framed condition is associated with large numbers and the negatively framed condition is associated with small numbers, the significant Framing × Ratee interaction supports Hypothesis 2, which states that the perceived performance difference between individuals is smaller when the objective behaviors are expressed with large absolute numbers than when they are expressed with small absolute numbers.

Furthermore, the three-way Framing × Ratee × Rating Dimensions interaction was not significant, $F(5, 375) = 1.66$, $MSE = .30$, $p > .1$, $\eta^2 = .02$, suggesting the significant Framing × Ratee interactions across all six rating dimensions were of comparable strength. Finally, there was a significant Ratee × Rating Dimensions interaction, $F(5, 375) = 14.65$, $MSE = .30$, $p < .0001$, $\eta^2 = .16$, indicating that the mean rating differences between Andy and William were greater for some dimensions than for others.

Discussion

This study generally supports Hypotheses 1 and 2. Two ratees were less discriminately rated when their objective performance was expressed with large numbers (e.g., 97% vs. 95%) than when their performance was expressed with small numbers (e.g., 3% vs. 5%). Furthermore, the same behavior is perceived to be more favorable when it is positively framed (e.g., presence rate of 97%) than when it is negatively framed (e.g., absence rate of 3%). The same patterns of results were also found in the overall ratings, suggesting that the perceived performance difference is not restricted to the dimensions presented with objective information. The overall
evaluations of the ratees, which were presumably derived from each performance dimension, were also affected by the S-shape function and the frame of reference.

In the next study, we attempt to disentangle a major confound inherent in the current research design. That is, number size and framing were confounded in this study. Currently, positively framed conditions were always presented with large numbers (e.g., presence rate of 97% or 95%) and negatively framed conditions were always presented with small numbers (e.g., absence rate of 3% or 5%). It is unclear whether it was the magnitude of the number or the framing per se that led to the change in the perceived performance difference in the two conditions. An alternative explanation of the present findings is that people perceived greater differences between Andy and William in the small (versus large) number condition because the performance data were presented in a negative (versus positive) way. Behavioral decision theory suggests that negatively framed events lead people to be more thorough in their analysis of information (Dunegan, 1993). As a consequence, they are less likely to overlook the differences between the two ratees in the negatively framed condition than they are in the positively framed condition. The differences in the rating discriminability may therefore likely be due to differential attention paid to the two framing conditions, rather than due to the magnitude of the number or the S-shaped subjective value function. We designed Study 2 to rule out this differential attention explanation by disentangling the confound between number sizes and framing (i.e., pairing large numbers with negative attributes and small numbers with positive attributes). Furthermore, the better performer (Andy) was always rated first in Study 1. Study 2 also helps rule out the confound of order as well.

Study 2

In this study, we also attempted to demonstrate that preference reversals could result from
switching the presentation format of the performance attributes (Hypothesis 3). Participants were told that they “are looking for computer programmers who can develop software for teaching purposes using a special computer language named KY.” They were further told that the two candidates (Andy and David), who were both new graduates, differed on two attributes, knowledge about KY and programming skill. Andy performed better than David in terms of knowledge about KY but worse than David in terms of programming skill. The information was presented in two versions (a David-favored versus an Andy-favored format) that differed in the presentation format of the performance attributes (see Table 3).

In the David-favored format, the attribute in which David was inferior to Andy (i.e., knowledge about KY) was presented with large numbers whereas the attribute in which David surpassed Andy (i.e., programming skill) was presented with small numbers. We expected that the mean overall rating for David would be higher than that for Andy. This is because the perceived difference between Andy and David on knowledge about KY was minimized (because the attribute was expressed with large numbers) and that on programming skill was exaggerated (because the attribute was expressed with small numbers). In the Andy-favored format, the attribute in which Andy was inferior to David (i.e., programming skill) was presented with large numbers whereas the attribute in which Andy outdid David (i.e., knowledge about KY) was presented with small numbers. Here, we expected to see a reversed pattern that the mean overall rating for Andy would be higher than that for David. This is because the perceived difference between Andy and David on Andy’s superior attribute is exaggerated and that on Andy’s inferior attribute is compressed. To recap Hypothesis 3 in more specific terms, we predict that

**Hypothesis 3:** Preference reversal will occur in the overall evaluation of David and Andy between two presentation formats. When the two performance attributes are expressed in a
David-favored format, raters will perceive David to be better than Andy whereas when the attributes are expressed in an Andy-favored format, raters will judge Andy to be better than David.

In this study, we recruited professional teachers to test Hypothesis 3. It should be noted that the professional teachers were familiar with performance evaluation and making between-individual comparisons because their typical duties include (1) school administration work, such as personnel selection, the task we used in Study 2; (2) regular evaluation of students’ performance (e.g., conduct, achievement, traits etc.); and (3) nominating students or fellow teachers for awards, scholarships, internships, or committee membership based on performance evaluations. The use professional teachers as the sample in this study complemented the use of students sample in Study 1, increasing the confidence that the results of our research could be generalized to people who are experienced in performance evaluation.

**Method**

**Participants, design, and evaluation task.** Participants were 40 professional teachers working in primary schools (56%) or secondary schools (44%). All respondents voluntarily participated in this study. Their average age was 34.7 years old, their average job tenure was 10.57 years, and 28% were males. There were no significant differences on these variables between the manipulated conditions.

Respondents were instructed that the purpose of their evaluations was to aid the school principal in making a hiring decision. Respondents were randomly given one of the two versions of a scenario in which the presentation format manipulation was embedded. In the David-favored format, knowledge about KY was indicated by “percentage of software written in KY that cannot be completed without consulting others” (ranging from 80% - 89%) and
programming skill was indicated by “percentage of software written in KY that can be used immediately without further debugging” (ranging from 3% - 9%). In the Andy-favored format, knowledge about KY was indicated by “percentage of software written in KY that can be completed without consulting others” (ranging from 11% - 20%) and programming skill was indicated by “percentage of software written in KY that cannot be used immediately without further debugging” (ranging from 91% - 97%). Information on the two candidates of the two attributes is provided in Table 3.

Respondents evaluated the candidates by indicating how satisfied they were with each candidate on a 7-point scale, with 1 = extremely dissatisfied and 7 = extremely satisfied. To counterbalance the potential effect of rating order, half of the respondents rated Andy before David and the other half rated David before Andy. Rating order did not have a significant main effect and did not interact with other variables on the ratings. This alleviates the concerns about order effects in Study 1.

After giving the performance ratings, participants were asked to indicate on a 5-point scale the attribute they regarded as more important when they made the performance judgment, with 1 = “Knowledge about KY is much more important than Programming skill”, 3 = “Knowledge about KY and Programming skill are equally important”, and 5 = “Programming Skill is much more important than Knowledge about KY”. We included this measure to check whether or not the perceived relative importance of the two performance attributes confounded our manipulations. The mean importance rating in the David-favored condition (3.14) was not significantly different from that in the Andy-favored condition (2.89), F (1, 38) = 1.23, MSE = .99, p > .1, η² = .03. We also carried out power analyses and found that the present study had a probability of .46 to detect the difference in the importance rating if the effect size was medium and had a probability of .8 to detect the difference if the effect size was large. Details of the
power analyses can be obtained from the first author on request. Thus, there was no support for the above confound.

Procedure. All aspects are the same as in Study 1, except that respondents were recruited through a professional teachers’ association.

Results

Figure 4 summarizes the results of this study. This figure clearly shows a cross over interaction pattern. The mean rating for David was higher than that for Andy in the David-favored condition while the opposite was true in the Andy-favored condition. A 2 (Format: Andy-favored vs. David-favored) × 2 (Candidate: Andy vs. David) ANOVA confirmed this observation. The focal Format × Candidate interaction was significant, $F(1, 38) = 15.81$, $\text{MSE} = .66$, $p < .01$, $\eta^2 = .30$. Planned comparisons revealed that in the David-favored condition, the mean rating for David (5.00) was significantly higher than that for Andy (4.24), $F(1, 38) = 9.28$, $\text{MSE} = .66$, $p < .005$, $\eta^2 = .44$. Conversely, in the Andy-favored condition, the mean rating for Andy (4.68) was significantly higher than that for David (4.00), $F(1, 38) = 6.77$, $\text{MSE} = .66$, $p < .05$, $\eta^2 = .21$. These findings therefore support Hypothesis 3.

Discussion

The results show that in situations involving a trade-off between performance attributes, preference can actually be reversed by switching the display format of objective performance. A
candidate is rated higher than another candidate when his/her superior attribute is expressed with a small number and his/her inferior attribute is expressed with a large number. These findings fit nicely with the S-shaped value function of prospect theory (see Figures 1 & 2) such that the difference in performance across ratees appears to be greater in small number conditions than in large number conditions.

The present findings do not support the differential attention explanation, which predicts that the difference in performance across ratees should be larger for negatively framed attributes than for positively framed attributes. In this study, small numbers were associated with positively framed attributes (e.g., 3% and 9% of software that can be used immediately without further debugging) and large numbers were associated with negatively framed attributes (e.g., 97% and 91% of software that cannot be used immediately without further debugging). The differential attention explanation would predict that the perceived difference between Andy and David should mainly be driven by their performance in the negatively framed attribute. Accordingly, David should be judged better than Andy in the Andy-favored format whereas the reverse would be true in the David-favored format. The present results contradict that prediction. Thus, this study not only replicates the findings of Study 1, but also rules out the differential attention explanation.

As pointed out by an anonymous reviewer, an interesting finding in Study 2 provides further support for prospect theory. In the Andy-favored condition, Andy was not rated as high as David was in the David-favored condition, although respondents in the two conditions indicated that the relative importance of the two performance attributes were comparable (e.g., 3.14 and 2.89 on a 5-point scale). According to prospect theory, the slope for 20% vs. 11% is slightly smaller than that for 3% vs. 9%. In turn, the superiority of Andy over David on “Knowledge about KY” in the Andy-favored condition (20% vs. 11%) was probably less impressive than the superiority of David over Andy on “Programming Skill” in the David-favored condition (9% vs. 3%).
difference in the marginal difference between Andy and David in their areas of excellence probably explains the pattern noted above

General Discussion

This paper has examined the role of between-individual comparisons on performance evaluation. We began with the hypothesis that performance evaluation is affected not only by the discrepancy between observed behaviors and the rater’s internal standard of performance, but also by the perceived performance differences among ratees. We noticed, however, that prior conceptual developments of performance evaluation have focused primarily on the behavior-standard discrepancy, whereas much less is known about the between-individual comparison aspect. We then offered the argument that prospect theory from judgment and decision-making research may provide a useful conceptual framework to understand the mechanism of between-individual comparisons in performance evaluation. Finally, we tested our hypotheses with two studies.

In this paper, we offer a contribution to the performance evaluation literature by borrowing ideas from prospect theory. Prospect theory was originally proposed as a theory of choice (Kahneman & Tversky, 1979, see also Tversky & Kahneman, 1991). In a choice task, the emphasis is on the outcome of a comparison on the subjective values of each alternative, whereas in a rating task, the focus is on the subjective values of each alternative per se. Hence, the subjective value function of prospect theory provides a useful framework for our research on performance rating. Our studies generally show that the between-individual comparison mechanism resembles the core characteristics of prospect theory. Studies 1 and 2 found that rating differences between two ratees was higher when objective performance was associated with small numbers than when it was associated with large numbers, supporting the non-linear nature of the subjective value function. Study 1 also found that objective performance was
perceived to be more favorable when it was positively framed than when it was negatively framed, supporting the idea of frame of reference. Finally, Study 2 demonstrated that the non-linear nature of the value function could lead to preference reversals by changing the expression of the performance attributes. This study also ruled out the confound that it was the framing of the performance attributes (positive or negative) that changed the perceived performance difference across individuals. Rather, it supported our central argument that it is the magnitude of the number that leads to the change.

Our results and the literature on contrast effects (Wexley et al., 1972; Wexley et al., 1973) show that the performance rating of an individual is affected by the performance of other individuals and by the way that information is presented. In particular, because between-individual comparison is a common purpose of performance evaluation (Cleveland et al., 1989) and because performance evaluation is influenced by its purposes (DeNisi, 1996; Murphy & Cleveland, 1995; Williams, DeNisi, Blencoe, & Cafferty, 1989; Zedeck & Cascio, 1982), understanding the mechanism underlying cross-individual comparisons in rating performance seems to be an important aspect of performance evaluation and deserves further research.

The present study also shows that knowledge from the field of judgment and decision-making (JDM) can contribute to understanding issues that are of primary concern in industrial and organizational psychology (I/O). Highhouse (1998) notes that although many topics within the domain of JDM are relevant to I/O, due to the differences in the dominant traditions in these two fields, there has been a gap between them such that many topics of JDM have been underrepresented in the I/O literature. In particular, Highhouse (1998) highlights that information presentation similar to the manipulation in the present study is an important factor that may affect judgment and decision-making in various domains in an organizational context. Responding to his call, we demonstrate that a conceptual framework from judgment and
decision-making research (e.g., prospect theory) could be useful for researchers to understand a topic that has been widely studied in I/O (e.g., performance evaluation). We second Highhouse’s suggestion that JDM has much to offer I/O and future research may consider more exchanges between these two fields.

There are several practical implications of our findings. The presentation of information in alternative forms may play a critical role in determining (1) the degree to which raters perceive a behavior as favorable and (2) the perceived performance difference between individuals. Study 1 shows that framing the same performance attribute in a positive versus negative way leads to more favorable performance rating. Across the two studies, presenting objective information with small numbers can lead to better discriminability on the attribute. For example, the perceived performance difference in attendance is likely to be more salient when it is presented as a small number (e.g., in terms of absence rate) than when it is presented as a large number (e.g., in terms of presence rate). This suggests that whether to present an attribute with a large number or a small number is not just a matter of convenience or convention. In the case with performance presented with large numbers, one may see the two individuals as comparable; in the other case with performance presented with small numbers, one may see a huge difference between the two people. The present findings suggest that the usefulness of an attribute in differentiating across individuals could be enhanced by presenting the attribute with small numbers.

Performance evaluation may involve multiple attributes. We agree that it may not always be desirable to amplify the differences across individuals especially on less important attributes. However, we do believe a better discriminability on core competencies and behaviors can ensure that managers compare individuals based on those key attributes. Criteria on which individuals are perceived as different from one another are more influential in the evaluation process than are those on which the individuals are perceived as similar. Thus, one way to minimize managers’
(un)intentional distortion on the relative weights of various criteria is to present the performance information using the appropriate format. For example, our findings suggest that raters would naturally perceive a bigger difference between candidates on a criterion when it is expressed with small numbers than when it is expressed with large numbers. In this way, we increase the likelihood that raters would differentiate the candidates on the basis of that particular attribute rather than on other information.

In sum, in situations when there is a mismatch in the presentation format and on the relative importance of the attributes, the quality of the evaluation may become questionable. For example, if performance on less important attributes is presented with small numbers while that on core attributes is presented with large numbers, a rater may perceive ratees as comparable in the key criteria while differentiating them based on the less important attributes.

The present findings also indicate that information display could alter the rating accuracy in terms of differential elevation, particularly when there are ratees with trade-offs between the performance attributes. As shown in Study 2, preference reversal was found such that one candidate was preferred in one presentation format, whereas the other one was preferred in another presentation format. This implies that opposing personnel decisions (e.g., selection or promotion) could be yielded when the same information is displayed differently. As between individual comparison is one of the major purposes of performance evaluation (Cleveland et al., 1989), accuracy in terms of differential elevation is very important in the organizational context (Murphy & Balzer, 1989; Murphy & Cleveland, 1995). Our findings demonstrate the potential threat of differential elevation if the information display format is not carefully chosen. One possible solution is to present both the negatively and positively framed information to raters, though the effectiveness of this suggestion should be tested in future studies.

Limitations and future directions
Our studies were highly controlled experiments that isolated other factors that occur in the real-world settings. This control increases the overall internal validity of the research and this is particularly important at the initial stage of understanding a phenomenon in performance evaluation (DeNisi, 1996; Murphy & Cleveland, 1995). Nevertheless, this could limit the generalizability of the findings to real-world settings in several ways.

The purposes and consequences of the ratings on the performance evaluation in the present study may not be the same as those in the field. In the field, raters often realize that the ratings on a performance evaluation will be used for administrative and/or developmental purposes. These purposes are likely to play an important role in determining the processes of performance evaluation (DeNisi, 1996; Murphy & Cleveland, 1995; Williams et al., 1985).

Between-individual comparison is one of the common purposes of performance evaluation (Cleveland et al., 1989) but the scenario method used in our research might make this purpose less salient. It is possible that the influences of between-individual comparisons on performance evaluation are more pronounced in the field than in our research.

There are purposes other than between-individual comparisons in the field. For example, managers may achieve political purposes by deliberately manipulating ratings to enhance or protect their self-interests (Cleveland et al., 1989; Longenecker, Sims, & Gioia, 1987). Murphy and Cleveland (1995) noted that “…the extent to which politics is operating in ratings vary considerably, depending on political culture of the organization – in particular, whether performance appraisal is taken seriously.” (p. 105; see also Longenecker et al., 1987 for a similar argument). We speculate that in organizations with a strong political culture, political biases might moderate the perceptual biases due to number sizes in two different ways. On the one hand, we may expect that political concerns are so dominant so that the impacts of presentation format of performance on evaluations may be overwhelmed by political biases. On the other
hand, it is also possible that managers reinforce perceptual biases by deliberately manipulating
the number sizes of performance information in a way consistent with their political interests.
Thus, how and under what conditions political biases overwhelm or reinforce perceptual biases
found in the present study are interesting issues for future research.

Thus far, we have argued that between-individual comparison follows the S-shaped function
of prospect theory. However, a broader prediction is that prospect theory may also be applicable
when comparing the performance within an individual at two time points (e.g., the perceived
improvement of a ratee over a time period) as well as when comparing an individual’s
performance to an absolute standard (e.g., the perceived absolute performance of a ratee). Both
tasks involve comparisons between two objective performances, i.e., performance at time 1
versus time 2 and actual performance versus an absolute standard. Presenting performance
information with small numbers may increase the perceived differences resulting from these
comparisons. For example, the improvement in accuracy expressed as a drop in error rates from
8% to 4% appears to be greater than that expressed as an increase in accuracy rates from 92% to
96%. Similarly, when 5% of error rates or 95% of accuracy rates are defined as the absolute
standard representing “acceptable performance”, it seems that an error rate of 8% (vs. 5%)
appears to be less acceptable than an accuracy rate of 92% (vs. 95%). We are currently
conducting studies to test these predictions.

Conclusion

Understanding how performance evaluation is influenced by between-individual
comparisons is critical to a better design of performance evaluation in rating tasks. Perceptions
of performance differences are often based on nuances in how information is presented. Here,
we have argued that prospect theory is an appropriate conceptual framework to describe the
between-individual comparison aspect of performance evaluation. In line with the core
predictions derived from this theory, our research suggests that the discriminability between ratees may vary depending on whether information about their performance is presented with large numbers or with small numbers.
References


Dunegan, K. J. (1993). Framing, cognitive modes, and image theory: Toward an understanding of
APA IN PRESS ARTICLE: If you plan to quote this manuscript, be sure to check the actual wording of the final published article.


Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. Organizational Behavior and Human Decision Processes, 76, 149-188.


APA IN PRESS ARTICLE: If you plan to quote this manuscript, be sure to check the actual wording of the final published article.

In J. B. Pryor & G. D. Reeder (Eds.), *The social psychology of HIV infection* (pp. 5-38). Hillsdale, NJ: Lawrence Erlbaum.


APA IN PRESS ARTICLE: If you plan to quote this manuscript, be sure to check the actual wording of the final published article.

Table 1. Presenting the same information in two different ways that yield preference reversal.

<table>
<thead>
<tr>
<th>Framing</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Andy</td>
<td>David</td>
</tr>
<tr>
<td>Programming Knowledge</td>
<td>2%</td>
<td>5%</td>
</tr>
<tr>
<td>(Failure rates)⁹</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic Record</td>
<td>B-</td>
<td>A</td>
</tr>
<tr>
<td>(Success rates)ᵇ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: ⁹ Failure rates = percentage of the written software that cannot be used; ᵇ Success rates = percentage of the written software that can be used
Table 2. Performance attributes by two framing conditions for Study 1.

<table>
<thead>
<tr>
<th>Framing</th>
<th>Negative direction</th>
<th>Positive direction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Andy</td>
<td>David</td>
</tr>
<tr>
<td>Attendance</td>
<td>3%</td>
<td>8%</td>
</tr>
<tr>
<td>Punctuality</td>
<td>13%</td>
<td>15%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>Completeness</td>
<td>7%</td>
<td>10%</td>
</tr>
<tr>
<td>Success</td>
<td>12%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Note.  

a In this condition, attendance = absence rate; punctuality = rate of being late; accuracy = error rate; completeness = incompleteness rate; success = customer-termination rate.  

b In this condition, attendance = presence rate; punctuality = rate of being on time; accuracy = accurate rate; completeness = completeness rate; success = customer retention rate.
Table 3. Performance information on David and Andy in Study 2.

<table>
<thead>
<tr>
<th>Presentation Format</th>
<th>David-favored format</th>
<th>Andy-favored Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge about KY</td>
<td>80% 89%</td>
<td>20% 11%</td>
</tr>
<tr>
<td>KY a</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Programming skill c</td>
<td>3% 9%</td>
<td>97% 91%</td>
</tr>
<tr>
<td>KY b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: a Knowledge about KY was expressed as “percentage of software written in KY that cannot be completed without consulting others” in the David-favored condition.

b Knowledge about KY was expressed as “percentage of software written in KY that can be completed without consulting others” in the Andy-favored condition.

c Programming skill was expressed as “percentage of software written in KY that can be used immediately without further debugging” in the David-favored condition.

d Programming skill was expressed as “percentage of software written in KY that cannot be used immediately without further debugging” in the Andy-favored condition.
Figure Captions

Figure 1. The subjective value function of prospect theory. Adapted from “Prospect Theory: An Analysis of Decision Under Risk,” by D. Kahneman and A. Tversky, 1979, *Econometrica*, 47, pp. 263-291. Copyright 1979 by The Econometric Society. Adapted with permission.

Figure 2. The subjective value function in the performance evaluation context. This function is drawn from the subjective value function of prospect theory. The figure illustrates how presenting information with large numbers or small numbers may influence the perceived difference between two performance instances. In the positive domain, the difference between 93% and 97% appears to be small because the two numbers are on the flat part of the value function. However, the difference between 7% and 3% in the negative domain appears to be large because the two numbers are on the steep part of the value function.

Figure 3. Results of Study 1. Performance evaluation as a function of Framing (+ve vs. –ve), Rating Dimensions (attendance, punctuality, accuracy, completeness, success, and overall), and Ratee (Andy vs. William).

Figure 4. Results of Study 2. Performance evaluation as a function of Format (Andy-favored vs. David-favored) and Candidate (Andy vs. David).
Figure 1
Figure 2

APA IN PRESS ARTICLE: If you plan to quote this manuscript, be sure to check the actual wording of the final published article.
Figure 3
Figure 4.

- **David-favored Format**: Andy: 4.6, David: 3.8
- **Andy-favored Format**: Andy: 5.0, David: 4.4

The graph shows a cross-over interaction where Andy prefers the Andy-favored format, while David prefers the David-favored format.