3-2007

Evaluating Dynamic Performance: The Influence of Salient Gestalt Characteristics on Performance Ratings

Jochen REB
Singapore Management University, jochenreb@smu.edu.sg

Russell CROPANZANO
University of Arizona
DOI: https://doi.org/10.1037/0021-9010.92.2.490.

RUNNING HEAD: Evaluating Dynamic Performance

Evaluating Dynamic Performance:
The Influence of Salient Gestalt Characteristics on Performance Ratings

Jochen Reb

Singapore Management University

Russell Cropanzano

University of Arizona

AUTHOR NOTE

An earlier version of this article was presented at the Academy of Management annual meeting, Honolulu, Hawaii, August 2005. We thank Gary Greguras for his helpful comments on earlier drafts of this article. Correspondence concerning this article should be addressed to Jochen Reb, Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899. E-mail: jreb@smu.edu.sg.

ABSTRACT

It is well recognized that performance changes over time. However, the effect of these changes on overall assessments of performance is largely unknown. In a laboratory experiment, we examined the influence of salient Gestalt characteristics of a dynamic performance profile on supervisory ratings. We manipulated performance trend (flat, linear-improving, linear-deteriorating, U-shaped, and ∩-shaped), performance variation (small, large), and performance mean (negative, zero, positive) within-subjects, and display format (graphic, tabular) between-subjects. Participants received and evaluated information about the weekly performance of different employees over a simulated 26-week period. Results showed strong main effects on performance ratings of both performance mean and performance trend, as well as interactions with display format. Theoretical and practical implications of the results are discussed.

KEYWORDS: Dynamic Performance, Gestalt Characteristics, Performance Evaluation, Performance Profile, Performance Ratings

EVALUATING DYNAMIC PERFORMANCE: THE INFLUENCE OF SALIENT GESTALT CHARACTERISTICS ON PERFORMANCE RATINGS

Performance evaluations have long been an important method for improving workplace effectiveness. These appraisals serve a variety of important purposes, such as identifying individuals for promotion, providing developmental feedback, underscoring training needs, and assigning merit pay (Cardy & Dobbins, 1994). Research has established a number of factors that affect performance ratings and can introduce imprecision and bias (Heslin, Latham, & VandeWalle, 2005; Landy & Farr, 1980), such as managers' political motives (Longnecker, Gioia, & Sims, 1987), appraisees' impression management (Frink & Ferris, 1998), the quality of the supervisor-subordinate relationship (Tepper, Uhl-Bien, Kohut, Rogelberg, Lockhart, & Ensley, 2006), the nature of the task to be rated (Lee, 1985), the purpose of the rating (Scullen, Mount, & Judge, 2003), and the social context in which the ratings are conducted (Levy & Williams, 2004).

Despite all of the important work that has examined various factors that affect performance appraisals, an important aspect has been relatively neglected: the nature of performance itself. Specifically, research on "dynamic criteria" has shown that worker performance can be *dynamic,* or subject to changes over time (e.g., Deadrick & Madigan, 1990; Ghiselli & Haire, 1960; Hoffman, Jacobs, & Gerras, 1992; Hulin, Henry, & Noon, 1990; Thoresen, Bradley, Bliese, & Thoresen, 2004). As illustrated by the example profile depicted in Figure 1a, a dynamic performance profile has a certain (a) performance mean (b) performance variation, and (c) performance trend.

Profiles of dynamic performance can reflect both longer-term and shorter-term changes in individual behavior. Longer-term changes, sometimes described as *performance trends*, may

increase or decrease *mean performance* over an extended period. These long-term trends can be due to changes in employee skills, knowledge, or experiences (e.g., Deadrick, Bennett, & Russell, 1997; Kanfer & Ackerman, 1989; Schmidt & Hunter, 1992; Schmidt, Hunter, & Outerbridge, 1986; 1988; Quiñones, Ford, & Teachout, 1995). For example, when a job is complex, performance tends to improve as experience increases (Sturman, 2003). In addition to these more stable changes, shorter-term fluctuations are also common. These may be described as *performance variation* around a constant mean. For example, one's immediate affective state can impact performance (e.g., Beal, Weiss, Barros, & MacDaniel, 2005; Cropanzano, Weiss, Hale, & Reb, 2003; Weiss & Cropanzano, 1996).

Despite the progress made on understanding the factors that influence fluctuations in performance, many questions remain (Austin & Villanova, 1992; Barrett, Caldwell, Alexander, 1985; Sturman Cheramie, & Cashen, 2005). One interesting question concerns how changes in performance over time influence overall performance ratings for that specific time period. In his work on performance distribution assessment, Kane (1986; 1996; for empirical tests, see e.g. Deadrick & Gardner, 1997; Jako & Murphy, 1990; Kane, 2000) developed perhaps the most sophisticated assessment method that explicitly acknowledges that performance is dynamic. While this work recognizes the potential influence of *performance variation* on evaluations, it pays less attention to *performance trends*, or the temporal unfolding of performance changes. However, several studies have provided evidence for an effect of performance trend on performance ratings (DeNisi & Stevens, 1981; Steiner & Rain, 1989; however, see also Scott & Hamner, 1975). For example, Steiner and Rain found that when performance was improving (a linear upward trend) ratings where higher than when performance was deteriorating (a linear downward trend) even when average performance was identical.

In perhaps the most comprehensive study to date, DeNisi and Stevens (1981) examined evaluations of eight different performance profiles that contained monthly sales figures over a six-month period. To generate these distributions, the authors first crossed three levels of performance mean (high, average, low) with two levels of performance variation (stable and variable) to yield six cells. To manipulate performance trend, one additional condition had an average performance mean and a linear-deteriorating trend and a last condition had the same performance mean but a linear-improving trend. They found that mean performance predicted ratings such that high performance was rated significantly better than average and low performance, whereas the latter did not differ significantly. Further, performance variation had no significant effect on performance ratings. Finally, results showed that improving performance was rated significantly better than deteriorating performance even though both performance profiles had the same mean performance.

This research suggests that when individuals rate job performance over an extended interval they do not simply add or average the performances at different points in time. They are sometimes influenced by additional factors, such as whether performance improves or deteriorates over the observation period. This finding is important because of the significant consequences that come with many performance evaluations. For example, an employee whose performance trend was deteriorating during the period shortly before the evaluation may be terminated even though the deterioration may have been due short-term environmental factors unrelated to the employee's effectiveness.

However, more research is needed. For one thing, a profile has at least three core features – performance mean, variation, and directional trend. None of the work to date has simultaneously manipulated all three of these attributes in a fully factorial design. However, past

research suggests that performance varies across all three dimensions (Deadrick et al, 1997; Harrison, Virick, & William, 1996; Sturman & Trevor, 2001; Sturman, 2003). Examining the relative effects of mean, variation, and trend on performance evaluations requires a full factorial design. In addition, no research has examined the potential moderating role of how the dynamic performance data is displayed. However, such a moderating effect could be potentially problematic as different supervisors may make their performance evaluations on the basis of performance profiles displayed in distinct formats. For example, if it turned out that performance trend had a stronger influence on performance ratings under a graphic display format (as argued in more detail below; see also Figures 1) then the same employee who showed a deteriorating performance would receive a worse rating from a supervisor using a graphic display of the performance profile than a supervisor using a tabular display. Furthermore, a challenge for addressing these issues is that to date no comprehensive theoretical framework has been proposed, much less tested, to account for ratings of dynamic performance. In this paper we will propose a theoretical model by borrowing from research on *Gestalt* characteristics.

Evaluation of Dynamic Performance and the Effect of Salient Gestalt Characteristics

When performance is dynamic, as it often is (e.g., Deadrick & Madigan, 1990; Hoffman, et al, 1992; Hulin et al, 1990; Thoresen et al, 2004), individuals who need to provide summary ratings, such as supervisors making semi-annual performance evaluations, are faced with a demanding cognitive task. They need to process a relatively large amount of information that can be arranged in various shapes (for an example, see Figure 1a). This information can consist of "objective" performance data, such as the sales revenues shown in Figure 1a, or in the subjective impressions derived from interacting with the subordinate, or a combination of both. Raters must then somehow integrate this information and represent it in a single score. As a result of limited

cognitive resources, individuals may lack the motivation or the capability to process larger amounts of data in an optimal manner. Indeed, most ratings forms do not even provide instructions as to how dynamic scores should be aggregated (see Kane, 1986, for an exception). Rather than carefully weigh each bit of available information, we propose that individuals will use simple, less effortful heuristics that allow them to process information in a holistic manner (Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982). As such, they should rely on salient Gestalt characteristics of the experience profile when making their judgments (Ariely & Carmon, 2000; Loewenstein & Prelec, 1993; Varey & Kahneman, 1992).

In this context, the term "Gestalt characteristics" is used in analogy to research on perception. This work has demonstrated that individuals tend not to form a simple composite of individual data point. Rather, they use salient structures of the perceived object. These attention-getting features help to organize the data points into a holistic perception (Koffka, 1935). By extension, the term "Gestalt" is used here as referring to the defining, salient features of a dynamic performance profile.

To our knowledge, work on Gestalt characteristics has not been applied to performance evaluation. Nevertheless, available work across a variety of judgment domains is consistent with our approach. For example, research has examined the influence of Gestalt characteristics on summary evaluations of extended experiences in such areas as monetary payments (Loewenstein & Sicherman, 1991), vacations (Loewenstein & Prelec, 1993), pain (Ariely, 1998; Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993), and medical outcomes and treatments (Chapman, 2000; Redelmeier & Kahneman, 1996), to name but a few. In the present study, we will focus on trend as a Gestalt characteristic that has been shown to be important.

Effect of Performance Trend

One Gestalt characteristics that has been repeatedly shown to affect summary evaluations is the *trend* of an extended profile (Ariely, 1998; Ariely & Carmon, 2000; Ariely & Zauberman, 2000; Chapman, 2000; Loewenstein & Prelec, 1993). Specifically, individuals tend to be more favorable towards an "upward" trend, as improvement is evaluated positively. At the same time, they tend to be less favorable towards a "downward" trend, as deterioration is evaluated negatively. Put differently, with mean level of performance held constant, an improving trend is evaluated more positively than a flat trend, and a flat trend more positively than a deteriorating trend (Ariely & Carmon, 2000). Based on this research we predict the following.

H1: Performance ratings will be higher for an improving trend than for a flat trend.

H2: Performance ratings will be lower for a deteriorating trend than for a flat trend.

Much less research has examined more complicated trends. In the present study, we also examine U-shaped and ∩-shaped trends. Ariely and Zauberman (2000) found the following ordering of summary evaluations of extended experiences (annoying sounds played over a period of time): deteriorating trend (i.e., sounds became more annoying over time) < ∩-shaped < U-shaped < improving trend. However, they did not include a flat trend condition. One might expect the effect of the U-shaped and ∩-shaped trends to be between the effects of the purely improving and deteriorating trends and a flat trend because the end of a profile tends to have a stronger impact than its beginning (Kahneman et al., 1993). Therefore, we hypothesize the following.

H3: Performance ratings for a U-shaped trend will be higher than for a flat trend and lower than for an improving trend.

H4: Performance ratings for an ∩-shaped trend will be lower than for a flat trend and higher than for a deteriorating trend.

Interaction of Dynamic Performance Characteristics with Display Format

Based on the notion of Gestalt characteristics, we propose that individuals rely on salient features of the stimulus in order to form a holistic judgment. The way information is presented will make certain features more salient and other features less so. Consistent with this, prior research has found that the format in which performance is displayed can impact ratings (e.g., Kulik & Ambrose, 1993; Wong & Kwong, 2005). In the present context, a presentation factor that might be influential is whether the display of information is graphic or tabular (see Figure 1a and 1b for examples of each). Tabular and graphic displays may differ at least in two ways with respect to which characteristics are salient to individuals with limited information processing capabilities. As such, display format (tabular versus graphic) might moderate the effect of dynamic performance characteristics, such as performance trend and mean.

First, in a graphic display performance trend tends to be more easily detectible than in a tabular display. This should be clear from comparing Figure 1a to Figure 1b. Notice that the direction of change is very easily detectible from examining the illustration, but less so from examining the table. Thus we hypothesize the following interaction.

H5: The effect of performance trend on performance ratings will be more pronounced in a graphic as compared to a tabular display.

On the other hand, tabular displays as the one displayed in Figure 1b may make it easier to detect differences in mean performance across ratees. For example, evaluators might simply estimate the ratio of the number of positive performances versus negative performances (easily recognizable by their minus sign in front of the number) to get a quick idea of whether performance tends to be more positive or negative (assuming the absence of outliers). Performance mean differences across different ratees may be harder to detect in a graphic display

because the visually more salient trend information will tend to direct limited attention away from information about the performance mean. Thus, we predict the following moderating effect of display format.

H6: The effect of performance mean on performance ratings will be more pronounced in a tabular as compared to a graphic display.

METHOD

Overview and Design

Participants assumed the role of a supervisor and completed semi-annual performance evaluations for their hypothetical subordinates. Each participant received a booklet containing information on 35 subordinates who were described as working in sales. The information consisted of 26 data points (one for each week over a span of 26 weeks) indicating the amount of dollars that a specific employee contributed to company revenues relative to the long-term average revenue contribution of employees in this company and this job. Raters were explicitly instructed to evaluate the *past* performance of the salespeople. We chose to focus on the evaluation of past performance because the argument has been made that mean performance should have the strongest, if not only effect on ratings of past performance, whereas other factors, such as performance trend or display format, should not affect performance (e.g., Scott & Hamner, 1975).

The experiment had a mixed 5 (performance trend: flat vs. improving vs. deteriorating vs. U-shaped vs. ∩-shaped) x 3 (performance mean: negative vs. zero vs. positive) x 2 (performance variation: small vs. large) x 2 (display format: tabular vs. graphic) design with the last factor being between-subjects. The manipulations were fully crossed, resulting in 30 within-subjects cells and two between-subjects conditions. Order was counterbalanced with one condition

reversing the order of the other. In addition, five employees were evaluated twice at different points during the study (excluded from the following analyses). Test-retest reliability for these five ratees, as assessed with correlation coefficients, was on average r =.66, which is within the 95% confidence interval of the estimate of test-retest reliability for subjective performance measures of high complexity jobs reported in Sturman et al's (2005) recent meta-analysis.

Participants

Sixty-four undergraduate business students at a large Southwestern public university participated in exchange for course credit. Sixty-eight percent were males. The average age of the sample was 21.3 years. Sixty-three percent indicated they had performed a self-evaluation as part of their job at least once (on average these participants had performed M = 4.91 self-evaluations, SD = 4.78), 43% indicated they had evaluated one or more subordinates at least once as part of their job (M = 5.67, SD = 4.78), and 61% had performed a peer evaluation as part of their job at least once (M = 6.76, SD = 8.59). Seventy-three percent said they had been evaluated by others at least once as part of their job (M = 5.14, SD = 5.62). Thus, the sample had non-negligible experience in evaluating others and being evaluated in a work context. Approximately 44% indicated they were employed at the time of data collection (average working hours per week, M = 28.7 hours; average tenure at current job, M = 22.3 months).

Procedure and Materials

All participants were given verbal and written instructions about the task. They received a package with an instruction sheet, the performance information of the 35 employees, and a short questionnaire to be filled out after the evaluations. They were seated in front of computers where they entered their evaluations using a keyboard. In the instructions, participants read that they were to assume the role of "Regional Supervisor" whose task was to review the performance

over the past 26 weeks and to give semi-annual performance evaluation to the junior-level sales personnel. They were informed that "these performance appraisals are used for personnel record keeping and to document your judgment of their overall performance over the pay period in question."

To further make sure that raters would focus on evaluating *past* performance as we intended they were told that, in addition to these performance ratings, they also needed to recommend qualified salespeople for promotion to the senior sales staff as their second task. At the end of the instructions, it was repeated again that raters had to make two judgments: first, "a rating of past performance based on your review of the 26 week appraisal period" (the rating we were interested in) and, second, "a promotion recommendation based on your prediction of future performance." Moreover, on top of the computer screen the program reminded raters in large, bright red letters of the rating purpose. Thus, for the first rating purpose it said "Evaluation of Past Performance" and for the second it said "Prediction of Future Performance."

Measure

Ratings were made on an 11-point scale ranging from –100 (labeled "worst performance") to +100 (labeled "best performance") on which supervisors rated the employee's performance over the past six months by selecting a number between –100 and 100 in steps of 20. A scale with a wide range was used to allow for a differentiation of the various performance profiles of the different ratees.

Manipulations

All three dynamic performance characteristics were manipulated within-subjects. Mean revenue contribution of an employee, relative to the company's long-term average for this job, was either negative –$1,800, zero ($0), or positive $1,800. Performance variation was either

relatively small (standard deviation = $200) or relatively large (standard deviation = $600). To create the different performance profiles in a first step a random number generator drew 26 (i.e., one for each week) numbers from a normal distribution with the respective mean and standard deviation required by the experimental condition. This implemented the performance mean and variation manipulations. In a second step, the performance trend manipulation was implemented. The trend was either flat, improving (in steps of +$150 per week), deteriorating (in steps of –$150 per week), U-shaped (in steps of –$300 for the first 13 weeks and +$300 for the second 13 weeks), or ∩-shaped (in steps of +$300 for the first 13 weeks and –$300 for the second 13 weeks). Thus, for all but the flat trend condition, the original distributions were transformed by adding or subtracting the appropriate values to implement the trend manipulation. To illustrate, the improving trend was implemented by subtracting $1875 from the data point generated in the first step for week 1, $1725 from the performance for week 2, and so on to week 26, for which $1875 was added to the original data point. Figure 1 gives as an example the profile that resulted from this process for the condition {zero mean; large variation; improving trend}.

Display format was manipulated between-subjects. Participants in the graphic display condition saw one figure for each employee on one page of their booklet, with the week given on the x-axis and the relative revenue contribution given on the y-axis (see Figure 1a for an example). Participants in the tabular condition saw one table for each employee on one page of their booklet (see Figure 1b for an example). The left column indicated the week and the right column the relative revenue contribution (i.e., performance) in that week.

Insert Figure 1a and Figure 1b around here

RESULTS

Our main analysis consisted in a 5 (trend) x 3 (mean) x 2 (variation) x 2 (display) mixed-measures ANOVA on performance ratings in which all experimental manipulations were treated as fixed effects and raters as random effects. All factors were manipulated within-subjects, except for display, which was manipulated between-subjects. For a tabulation of means, see Table 1; for a summary of ANOVA results and effects sizes, see Table 2.

Insert Table 1 around here

Hypotheses 1-4: The Effect of Performance Trend

The main effect for performance trend was significant, $F(4, 248) = 68.21$, $p < .001$, $\omega^2 = .05$ (see Figure 2). Pairwise comparisons were used to test Hypotheses 1-4. In the improving trend condition ($M = 25.05$) performance was evaluated significantly higher than in any other trend condition (all $p < .001$), consistent with Hypothesis 1. The difference between mean ratings in the improving trend and the deteriorating trend ($M = -16.98$) *given same mean performance* (i.e., same amount of revenue contributed to the company) is a rather striking 42 points (on a scale ranging from -100 to +100). The deteriorating trend produced significantly lower ratings than any other trend (all $p < .001$), consistent with Hypothesis 2. The U-shaped trend ($M = 12.19$) was right between the improving and flat ($M = 2.14$) trends and significantly different from both (both $p < .001$), consistent with Hypothesis 3. Performance ratings in the ∩-shaped trend condition ($M = -1.25$) were higher than for the deteriorating trend ($p < .001$), but not significantly lower than for the flat trend ($p = .11$), only partly consistent with Hypothesis 4.

In sum, performance trend had a substantial influence on the evaluation of subordinates. Moreover, the form of the effect largely followed our expectations. The only exception was the ∩-shaped trend, which produced higher ratings than expected. One explanation could be the influence of the peak of the performance distribution, which was highly positive, on overall ratings. This explanation is consistent with findings in other judgment domains showing strong effects of the peak of a distribution on summary evaluations (e.g., Varey & Kahneman, 1992).

Insert Figure 2 around here

Hypothesis 5: The Interaction of Performance Trend with Display Format

As predicted, ANOVA showed a significant interaction between display and trend, $F(4, 248) = 4.06$, $p < .01$, $\omega^2 = .003$. However, inspection of means (see Figure 2) suggests the form of the interaction is only partially consistent with Hypothesis 5. As expected, the graphic display led raters to evaluate the improving trend more positively ($M = 33.44$) than the tabular display ($M = 16.67$), $F(1, 62) = 9.18$, $p < .01$. However, the same was the case for all trends except the deteriorating trend, which was evaluated about equally in both display conditions ($M = -18.02$, graphic, $M = -15.94$, tabular), $F(1, 62) = .12$, *ns*. In other words, display condition moderated the influence of trend on ratings such that a graphic display led to more positive evaluations for all trends except the deteriorating trend.

Hypothesis 6: The Interaction of Performance Mean and Display Format

As expected, ANOVA revealed a significant interaction between display and mean performance, $F(2, 124) = 72.05$, $p < .001$, $\omega^2 = .05$. Inspection of means (See Figure 3) suggests that the nature of the interaction is consistent with Hypothesis 6. Specifically, positive mean

performance was rated more positively in the tabular condition (M = 70.31) than in the graphic condition (M = 48.19), $F(1, 62) = 21.90$, $p < .001$. In contrast, negative mean performance was rated more negatively when the display was tabular (M = -75.69) than when it was graphic (M = -29.06), $F(1, 62) = 61.71$, $p < .001$. Finally, when mean performance was zero, display format did not affect ratings (tabular, M = 1.63, graphic, M = 10.00), $F(1, 62) = 3.13$, *ns*.

Insert Figure 3 around here

Interestingly from a practical perspective, Figure 3 also suggests a substantial main effect of mean performance on performance ratings. ANOVA confirmed this impression by revealing a highly significant main effect for mean, $F(2, 124) = 756.95$, $p < .001$, $\omega^2 = .58$. Inspection of the means shows that the effect is almost perfectly linear, increasing from M = -52.38 for the negative mean, to M = 5.81 for the zero mean, to M = 59.25 for the positive mean. These results are interpretable as the effect of mean performance is only partially qualified by the interaction with display – high performers tend to get higher ratings, regardless of the display format. These findings are somewhat reassuring from an applied perspective. After all, from a normative perspective, it could be considered a minimum requirement that ratings of past performance are affected by average performance (DeNisi & Stevens, 1981; Scott & Hamner, 1975).

Ancillary Analysis: The Null Effect of Performance Variation

As we discussed in the Introduction, neither Scott and Hamner (1975) nor DeNisi and Stevens (1981) found an effect for performance variation. Replicating these earlier results, ANOVA showed no effect of performance variation on performance ratings, $F(1, 62) = .50$, *ns*, $\omega^2 = 0$. However, it is, of course, possible that in our study, as well as the earlier studies, the

manipulation of variation was too weak and ineffective in either being noticed by or attracting sufficient attention from raters.

Insert Table 2 around here

DISCUSSION

Building on previous research (e.g., DeNisi & Stevens, 1981; Scott & Hamner, 1975) we examined how raters evaluate dynamic performance. Research participants completed ratings of 26-week performance profiles of hypothetical salespeople. In this context, we manipulated within-subjects in a fully-crossed design the performance trend (flat, improving, deteriorating, U-shaped, and ∩-shaped), mean (negative, zero, and positive), and variation (small, large) of the ratees. We also manipulated between-subjects whether the performance profiles of employees were presented in graphs or in tables.

Based on the idea that individuals use salient Gestalt characteristics when integrating the information contained in dynamic profiles into summary evaluations, and that trend is such a salient characteristic (Ariely & Carmon, 2000), we predicted that performance trend would influence performance ratings. As expected our analyses revealed a substantial effect of performance trend on evaluations, with the pattern of the effect being consistent with findings in other judgment domains (e.g., Ariely, 1998; Chapman, 2000; Loewenstein & Prelec, 1993; Varey & Kahneman, 1992). Specifically, employees' past performance was rated more positively when it showed an improving trend rather than a flat trend (given same average performance). A flat trend, in turn, was rated higher than a deteriorating trend. As expected, the U-shaped trend produced ratings between the flat and improving trends. However, while the ∩-shaped trend was

rated higher than the deteriorating trend it was not rated significantly lower than the flat trend. As mentioned in the Results section, it is possible that the high peak in this condition ($3425 in week 13) was largely responsible for the result (cf. Varey & Kahneman, 1992).

Theoretical Contributions

This study, along with other research on dynamic performance (e.g., Deadrick, Bennett, & Russell, 1997; Hoffman, Jacobs, & Gerras, 1992; Sturman et al, 2005), is significant because it changes the way we think of an important psychological construct and its assessment. As such, it contributes to the literature on dynamic criteria, and, more generally, the "criterion problem" (e.g., Austin & Villanova, 1992; Borman, 1978; 1991; Deadrick & Madigan, 1990; Smith, 1976; Sturman et al, 2005). The "job performance" supervisors evaluate is not a single homogenous episode (DeNisi & Stevens, 1981; Kane, 1986). Rather, it is an unfolding procession of events whose character is shaped, to some non-trivial degree, by the way in which they unfold. The shape in which performance develops over time can have important consequences on how it is evaluated – over and above the mean performance level. An analogy might help. If one wanted to describe the depth of a calm lake, a simple average might be sufficient. However, to describe the tides in a harbor it would be useful to have an appreciation of they way in which the average depth changes lawfully at particular time intervals. Similarly, when evaluating performance it may not be enough to know only the mean performance level. Depending on the rating purpose, it could be worthwhile to know if the work of one employee is deteriorating while another is improving.

None of the preceding is to say that mean performance is unimportant. Indeed, our study found that participants distinguished mean performance differences regardless of presentational format (see Figure 3, for example) and that the main effect for mean performance accounted for

most variance among all examined effects (see Table 2 for ω^2 effect size estimates). Rather, our point is that mean performance is not the only attribute distinguishing performance profiles. Our participants were considering information in addition to the mean, as witnessed by the substantial effect of performance trend as well as the interactions between display format and mean performance, and display format and performance trend. This was true even though we explicitly asked raters to evaluate past performance. It is possible that raters engaged in a sort of *naïve extrapolation*, naturally incorporate into their evaluations their expectations of future states (Ariely & Carmon, 2003).

In view of all this, there is another contribution of our study. We do not simply point out a potential problem with performance evaluations. In addition, we have presented a theory that makes sense of the phenomenon. Based on the notion of Gestalt characteristics, we were able to identify aspects of the performance distribution, as well as its presentation, that impact ratings in predictable ways. Our results lend support to the idea that raters make use of salient characteristics when evaluating dynamic performance profiles. Future research should examine other Gestalt characteristics that were not considered here. Ariely and Carmon (2003) distinguish between dynamic Gestalt characteristics, such as trend, and static Gestalt characteristics, such as peak and end of the experience profile (cf. Fredrickson & Kahneman, 1993; Kahneman et al., 1993). The present research has focused on trend as a dynamic characteristic. Future research could examine additional dynamic characteristics of a performance profile such as skewness, but also static characteristics, such as positive and negative performance peaks. Of special interest might be the role of outliers in performance. For example, it could be that a downward outlier (i.e., a single extremely bad performance) can have an unduly strong effect on performance ratings.

Finally, by examining the moderating role of display format our study also began to examine conditions under which Gestalt characteristics of a performance profile exert a weaker or stronger effect. For example, if one wished to increase the impact of mean level performance, then a tabular display might be appropriate (see Hypothesis 6). Of course, our results should, however, not be interpreted as implying that the effect of mean will always be stronger in a tabular than in a graphic display format. Rather, we think that this result only illustrates the more general idea that different display formats make different aspects of a dynamic performance profile salient. Future research should examine more systematically how display formatting can increase or decrease the salience of certain characteristics of a dynamic performance profile. For example, we would expect that the effect of performance trend in a graphic display format might be less pronounced if the y-axis ranged from -20,000 to +20,000 instead of -6,000 to +6,000 as in the present study. Similarly, we would expect the effect of mean performance in a tabular display format to be less pronounced if all performance data were positive (or negative) and, therefore, raters could not use the proportion of minus signs to help assess overall performance.[1]

Practical Contributions

Following from these theoretical issues, our results suggest a practical concern for interpreting performance evaluations. Because an overall rating is likely influenced by the salient features of the distribution, over and above the mean performance level, then comparing two scores may be more difficult than previously expected. For example, a person with a lower mean performance but an upward trend could receive a higher rating than a coworker with a higher mean performance but a flat trend. To illustrate, we conducted some ancillary analyses. In so doing, we found several cases in which employees who contributed on average $1,800 less per week ($46,800 over the 26 week period) were rated just as good as their better-performing peers

with respect to their past accomplishments because of differences in performance trend (i.e., improving for the worse performers versus deteriorating for the better performers).

Thinking about the matter this way suggests a potential solution that is in need of additional research. Rating procedures could be devised for decomposing the judgments, having raters make several specific appraisals (for mean level, for trend etc.) rather than a single holistic evaluation. If the appropriate evaluation technologies could be developed, such procedures could have extensive practical value. Kane's (1986; 2000) work on performance distribution assessment has made important progress towards this goal. However, this approach does not consider the potentially valuable information contained in the performance trend. Incorporating the temporal sequence of the performance data into this approach could provide an additional step towards an adequate assessment of job performance as it unfolds over time.

Being able to characterize the performance distribution in this manner can afford managers specific information tailored to particular decision needs. For example, if one is attempting to award end-of-year bonus money, then mean performance might be most appropriate. If the bonus is a reward for past behavior, then it would seem reasonable to "partial out" the effects of other distributional characteristics. On the other hand, if one is attempting to project future performance, then trend data would seem especially valuable. Variation and trend information might also be valuable in providing developmental feedback and detecting problems. For example, an inverted-U might signal an employee who is undergoing a personal crisis or experiencing some sort of stress. Further, because performance trends have been found to be related to turnover (Harrison et al, 1996; Sturman & Trevor, 2001), trend information might be helpful in predicting and preventing turnover. The key idea is that each feature of the distribution could have distinct causes and, in principle, distinct solutions. If all three features of the

distribution are lumped together, then different informational cues are obfuscated. Future research should explore whether rating purpose interacts with these features to influence performance ratings.

Limitations and Future Research

As with any study, this study has several limitations. First, our raters engaged in evaluations of hypothetical employees. On the one side, this allowed us the control to manipulate different aspects of the performance profile orthogonally. On the other side, the incentives for accurate assessment were relatively low and the context was somewhat artificial. Future research should attempt to study the role of performance profile characteristics in real performance evaluations. Also, the raters themselves were undergraduate business students, not actual members of organizations routinely performing performance appraisals, even though many of our participants had non-negligible experience with evaluating the performance of self and others and had work experience. We conducted two additional ANOVAs to examine whether (1) employment status (whether or not the participant was employed at the time of the study), and (2) evaluation of others (whether or not the participant had ever formally evaluated at least one other person as part of his or her job) affected our results. The analyses showed no significant interactions of these two variables with any of the experimental manipulations. While this provides some additional credibility to the present sample, clearly, future research should examine whether our results hold for more experienced raters.

In our experiment, raters were presented at a single point in time with data from the last 26 weeks. This procedure has been commonly used in other research (e.g., Chapman, 2000; Loewenstein & Prelec, 1993). However, future research should let raters sample performance over a longer period of time. Obviously, in the laboratory a period of 26 weeks is not feasible,

but a performance profile could be sampled over a period of, for example, an hour. Moreover, in our study, performance profiles consisted of "objective" performance data (i.e., actual sales figures). A valuable extension to address these issues would be to examine in an organizational setting how a series of subjective evaluations that are based on a manager's experiences with a subordinate over a longer time period are integrated into an overall performance evaluation (e.g., in the context of a semi-annual performance review).

Another direction for future research is to examine whether our findings hold for other evaluation targets. Specifically, how do performance profile characteristics affect the evaluation of one's own performance? Most of the research in other judgment domains has actually been concerned with ratings of one's own experience. In contrast, our study was concerned with the evaluation of someone else's performance. It is possible that some of the differences in results between our study and past research are due to this factor. In this sense, our study contributes to research on the summary evaluation of past experiences by highlighting the potential moderating role of evaluation target (self versus other).

Finally, we have examined dynamic performance from the view of the rater. A different, yet complementary perspective would be from the ratee's point of view. Specifically, it is well known that appraisees use impression management strategies in an attempt to influence performance evaluations (Frink & Ferris, 1998). Future research should examine whether employees have knowledge of the effect of performance trend on appraisals and try to take this effect into account. Specifically, our results suggest that employees might be tempted to produce increasing performance trends, cumulating in maximum performance right towards the end of the appraisal period in an effort to receive more favorable evaluations.

Conclusion

Research on performance appraisals has largely neglected the evaluation of dynamic performance and the influence of characteristics of dynamic performance profiles on performance ratings. However, research in other domains has discovered that individuals tend to evaluate dynamic profiles extending over time in a holistic manner, relying on a few salient Gestalt characteristics to arrive at their summary evaluations. We have built on this research and demonstrated the role of one Gestalt characteristic, the performance trend, in shaping overall performance evaluations. Moreover, we have shown how features of the display format can affect ratings by making salient different aspects of the performance profile. The influence of such salient characteristics could potentially be problematic given a goal of accurate and fair performance appraisals. On the other side, explicitly considering the dynamic nature of performance might provide useful information for a variety of appraisal purposes, such as providing developmental feedback. Future research should examine in more detail the descriptive and normative issues involved in evaluating the dynamic performance of self and others.

ENDNOTES

1 We thank one of our reviewers for these two examples.

REFERENCES

Ariely, D. (1998). Combining experiences over time: The effects of duration, intensity changes and on-line measurements on retrospective pain evaluations. *Journal of Behavioral Decision Making, 11,* 19-45.

Ariely, D., & Carmon, Z. (2003). Summary assessment of experiences: The whole is more than the sum of its parts. In Loewenstein, G., Read, D., & Baumeister, R. (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice.* New York: Russell Sage Foundation.

Ariely, D., & Carmon, Z. (2000). Gestalt characteristics of experiences: The defining features of summarized events. *Journal of Behavioral Decision Making, 13,* 191-201.

Ariely, D., & Zauberman, G. (2000). On the making of an experience: The effects of breaking and combining experiences on their overall evaluation. *Journal of Behavioral Decision Making, 13,* 219-232.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77*, 836-874.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1985). The concept of dynamic criteria: A critical reanalysis. *Personnel Psychology, 38*, 41-56.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in performance ratings. *Journal of Applied Psychology, 63*, 135-144.

Beal, D. J., Weiss, H. M., Barros, E., & MacDaniel, S. M. (2005). An episodic process model of affective influences on performance. *Journal of Applied Psychology, 90,* 1054-1068.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp.

271–326). Palo Alto: Consulting Psychologists Press.

Cardy, R. L., & Dobbins, G. H. (1994). *Performance appraisal: Applied perspectives.* Cincinnati, OH: South-Western Publishing.

Chapman, G. (2000). Preferences for improving and declining sequences of health outcomes. *Journal of Behavioral Decision Making, 13,* 203-218.

Cropanzano, R., Weiss, H. M., Hale, J. M. S., Reb, J. (2003). The structure of affect: Reconsidering the relationship between negative and positive affectivity. *Journal of Management, 29,* 831-857.

Deadrick, D. L., Bennett, N., & Russell, C. J. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management, 23,* 745-757.

Deadrick, D. L., & Gardner, D. G. (1997). Distributional ratings of performance levels and variability. *Group & Organization Management, 22*, 317-342.

Deadrick, D. L. & Madigan, R. J. (1990). Dynamic criteria revisited: A longitudinal study of performance stability. *Personnel Psychology*, 44, 717-744.

DeNisi, A., & Stevens, G. E. (1981). Profiles of performance, performance evaluations, and personnel decisions. *Academy of Management Journal, 24,* 592-602.

Frink, D. D., Ferris, G. R. (1998). Accountability, impression management, and goal setting in the performance evaluation process. *Human Relations, 51*, 1259-1283.

Fletcher, C. (2001). Performance appraisal and management: The developing research agenda. *Journal of Occupational and Organizational Psychology, 74,* 473-487.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65,* 45-55.

Ghiselli, E. E., & Haire, M. (1960). The validation of selection tests in the light of dynamic

character of criteria. *Personnel Psychology, 13,* 225-231.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, U.K.: Cambridge University Press.

Harrison, D. A., Virick, M., & William, S. (1996). Working without a net: Time, performance, and turnover under maximally contingent rewards. *Journal of Applied Psychology, 81*, 331-345.

Heslin, P. A., Latham, G. P., & VandeWalle, D. (2005). The effects of implicit person theory on performance appraisal. *Journal of Applied Psychology, 90,* 842-856.

Hoffman, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77,* 185-195.

Hulin, C. L., Henry, R. & Noon, S. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relations. *Psychological Bulletin*, 107, 328-340.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54,* 321-368.

Jako, R. A., & Murphy, K. R. (1990). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. *Journal of Applied Psychology, 75*, 500-505.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4,* 401-405.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, U.K.: Cambridge University Press.

Kane, J. S. (1986). Performance distribution assessment. In R. A. Berk (Ed.), *Performance*

assessment: Methods and applications (pp. 237-273). Baltimore: Johns Hopkins University Press.

Kane, J. S. (1996). The conceptualization and representation of total performance effectiveness. *Human Resource Management Review, 6,* 123–145.

Kane, J. S. (2000). Accuracy and its determinants in distributional assessment. *Human Performance, 13*, 47-84.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 75*, 657-690.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt Brace.

Kulik, C. T., & Ambrose, M. L. (1993). Category-based and feature-based process in performance appraisal: Integrating visual and computerized sources of performance data. *Journal of Applied Psychology, 78,* 821-830.

Landy, F.J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107.

Lee, C. (1985). Increasing performance appraisal effectiveness: Matching task types, appraisal process, and rater training. *Academy of Management Review, 10,* 322-331.

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

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review, 100,* 91-108.

Loewenstein, G., & Sicherman, N. (1991). Do workers prefer increasing wage profiles? *Journal of Labor Economics, 9*, 67-84.

Longnecker, C. O., Gioia, D. A., Sims, H. P. (1987). Behind the mask: The politics of employee

appraisal. *Academy of Management Executive, 1,* 183-193.

Quiñones, M.A., Ford, J.K., & Teachout, M.S. (1995). The relationship between work experience and job performance: A conceptual and meta-analytic review. *Personnel Psychology*, 48, 887-910.

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*, 3-8.

Schmidt, F. L., & Hunter, J. E. (1992). Development of causal models of process determining job performance. *Current Directions in Psychological Science, 1,* 89-92.

Schmidt, F. L, Hunter, J. E., & Outerbridge, A. N. (1986). The impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432-439.

Schmidt, F. L, Hunter, J. E., & Outerbridge, A. N. (1988). The joint relation of experience and ability with job performance: A test of three hypotheses. *Journal of Applied Psychology, 73,* 46-57.

Scott, W. E. Jr., & Hamner, W. C. (1975). The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. *Organizational Behavior and Human Performance, 14,* 360-370.

Scullen, S. E., Mount, M. K., & Judge, T. A. (2003). Evidence of the construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology, 88,* 50-66.

Smith, P. C. (1976). Behavior, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–775). Chicago: Rand McNally.

Steiner, D. D., & Rain, J. S. (1989). Immediate and delayed primacy and recency effects in

performance evaluation. *Journal of Applied Psychology, 74,* 13-142.

Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management, 29*, 609-640.

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test--retest reliability of employee job performance ratings. *Journal of Applied Psychology, 90*, 269-283.

Sturman, M. C., & Trevor, C. O. (2001). The implications of linking the dynamic performance and turnover literatures. *Journal of Applied Psychology, 86*, 684-696.

Tepper, B. J., Uhl-Bien, M., Kohut, G. F., Rogelberg, S. G., Lockhart, D. E., & Ensley, M. D. (2006). Subodinates' resistance and managers' evaluations of subordinates' performance. *Journal of Management, 32,* 185-209.

Thoresen, C. J., Bradley, J. C., Bliese, P. D., & Thoresen, J. D. (2004). The big five personality traits and individual job performance growth trajectories in maintenance and individual stages. *Journal of Applied Psychology, 89,* 835-853.

Varey, C. A., & Kahneman, D. (1992). Experiences extended across time: evaluation f moments and episodes. *Journal of Behavioral Decision Making, 5,* 169-185.

Weiss, H. M., & Cropanzano, R. (1996). An affective events approach to job satisfaction. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 18, pp. 1-74). Greenwich, CT: JAI Press.

Wong, K. F. E., & Kwong, J. Y. Y. (2005). Between-individual comparisons in performance evaluations: A perspective from prospect theory. *Journal of Applied Psychology, 90,* 284-294.

Table 1

Performance Ratings as a Function of Performance Trend, Performance Mean, and Display Condition

Performance Mean	Performance Trend					
	Deteriorating	∩-shaped	Flat	U-shaped	Improving	Total
			Graphic Display			
Negative						
M	-17.19	9.38	3.75	21.25	32.81	10.00
SD	39.66	29.81	39.82	21.93	33.88	37.51
Zero						
M	21.25	40.31	57.19	54.38	65.31	47.69
SD	37.35	29.28	29.57	31.36	24.23	34.18
Positive						
M	-58.13	-37.50	-33.44	-18.44	2.19	-29.06
SD	34.68	37.46	40.95	39.57	40.96	43.50
Total						
M	-18.02	4.06	9.17	19.06	33.44	9.54
SD	49.31	45.47	52.49	43.48	42.35	49.69
			Tabular Display			
Negative						
M	-87.19	-80.31	-83.13	-71.56	-56.25	-75.69
SD	17.59	23.50	21.74	29.34	45.06	30.91
Zero						
M	-8.75	-1.88	-15.94	9.06	25.63	1.63

SD	39.82	30.60	39.87	33.98	38.29	39.27
Positive						
M	48.13	62.50	84.38	78.44	80.63	70.81
SD	38.66	27.02	20.92	19.29	20.77	29.49
Total						
M	-15.94	-6.56	-4.90	5.31	16.67	-1.08
SD	64.90	64.50	74.74	67.56	66.89	68.59
Across Display Conditions and Performance Means						
M	-16.98	-1.25	2.14	12.19	25.05	4.23
SD	57.57	55.98	64.88	57.15	56.54	60.11

Table 2

Analysis of Variance Results for Effects Reported in Results Section

Source	*df*	*SS*	*MS*	*F*	ω^2
Performance Trend	2	376753	94188	68.21**	.05
Performance Trend x Display	2	22415	5604	4.06*	.003
Error (Trend)	124	342432	1381		
Performance Mean	4	3989652	1994826	756.95**	.58
Performance Mean x Display	4	379727	189863	72.05**	.05
Error (Mean)	248				
Performance Variation	1	441	441	.50	.00
Error (Performance Variation)	62	54312	876		

*$p < .01$. **$p < .001$.

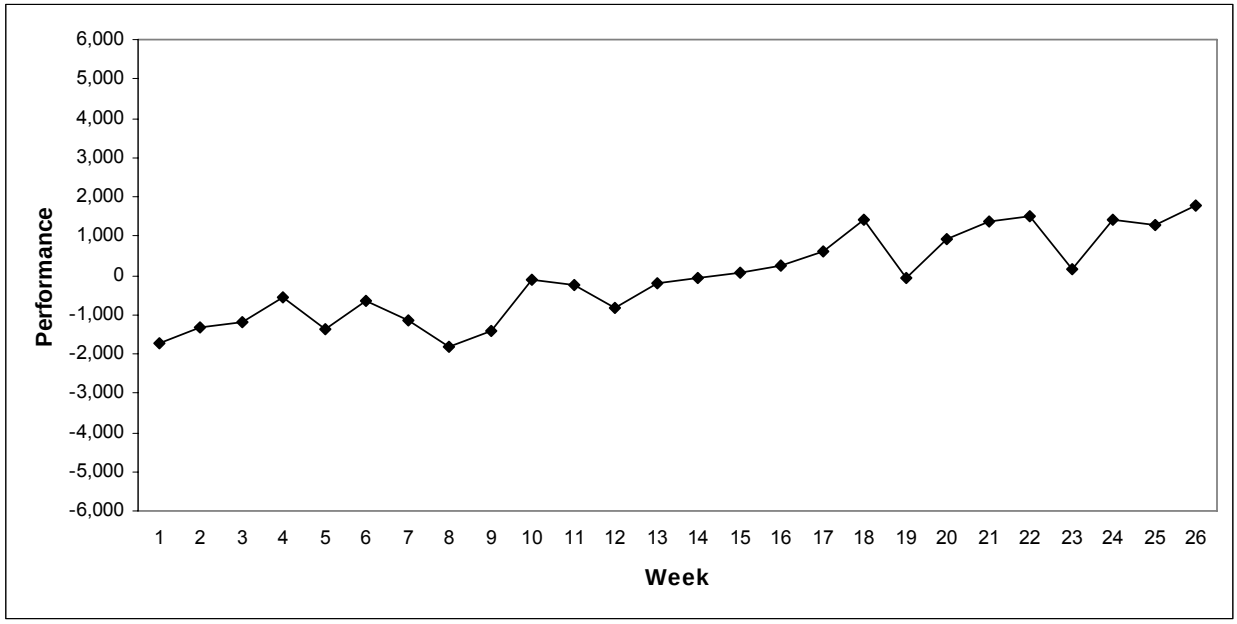
FIGURE CAPTIONS

Figure 1a. Example of Graphic Display of Dynamic Performance Profile

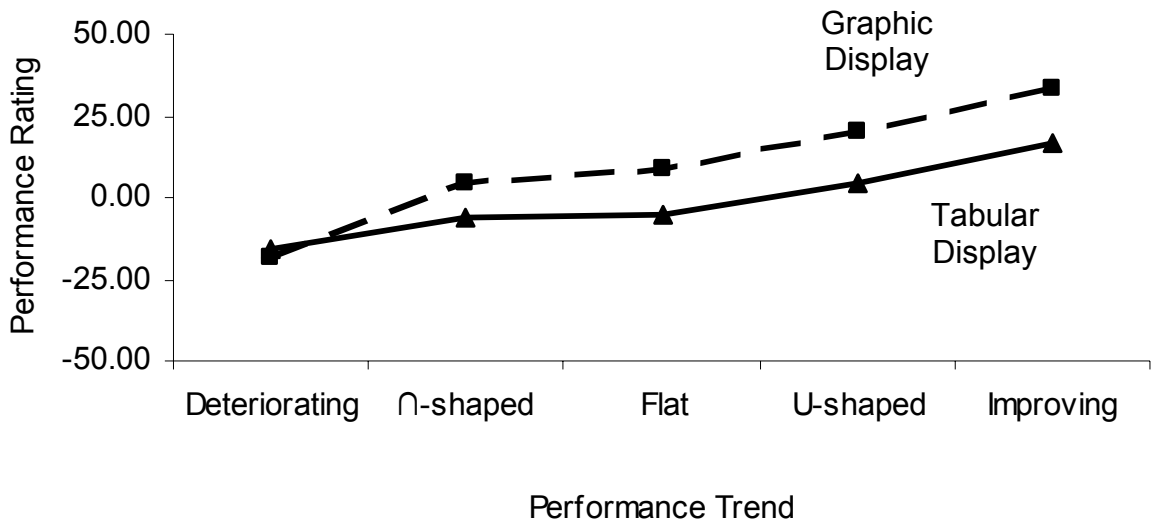
Figure 1b. Example of Tabular Display of Dynamic Performance Profile

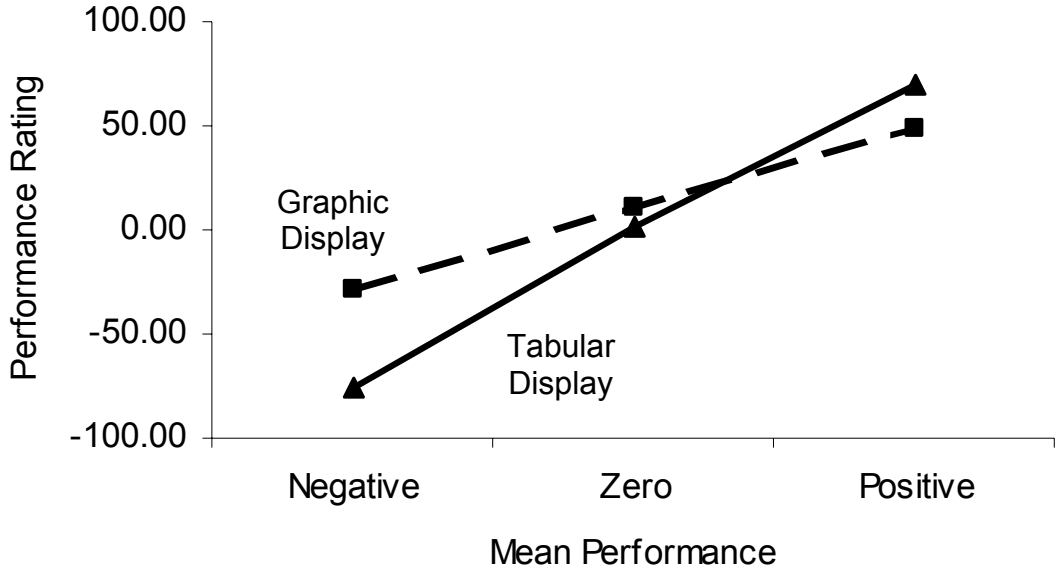
Figure 2. Effect of Performance Trend on Performance Ratings Depending on Display Condition

Figure 3. Effect of Performance Mean on Performance Ratings Depending on Display Condition


6,000
5,000
4,000
3,000
2,000
1,000
0
-1,000
-2,000
-3,000
-4,000
-5,000
-6,000
Performance
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Week

Week	***Revenue Contribution (in US dollars)***
Week 1:	**-1,727**
Week 2:	**-1,314**
Week 3:	**-1,202**
…	…
Week 24:	**1,403**
Week 25:	**1,287**
Week 26:	**1,766**


Graphic Display
Tabular Display
50.00
25.00
0.00
-25.00
-50.00
Performance Rating
Deteriorating
∩-shaped
Flat
U-shaped
Improving
Performance Trend


100.00
50.00
0.00
-50.00
-100.00
Performance Rating
Graphic Display
Tabular Display
Negative
Zero
Positive
Mean Performance