

Testing Assumptions: Can Performance Rating Feedback Result in Objective Performance Improvements?

by

Jennifer Gonder - Farmingdale State College

William Metlay - Hofstra University

Terri Shapiro - Hofstra University

Abstract

Performance appraisal is intended to be an instrument for performance improvement and the use of ratings is based on the assumption that rating feedback will have an impact on objective performance outcomes. Yet, most studies measure improvement as changes in performance ratings over time and there is limited empirical evidence to support this assumption. The present experiment was designed to address this gap in the literature by directly testing the effect of rating feedback on objective performance. Groups completed two problem solving tasks, receiving rating feedback on several team performance dimensions after the first task. Improvement in ratings and objective performance scores on the second task was measured to evaluate the impact of the feedback. Results replicated past research, demonstrating that rating feedback leads to improvement in subsequent ratings. Further, this experiment provides empirical evidence in support of the assumption that rating feedback can result in objective performance improvement.

Keywords: performance appraisal, ratings, feedback, performance improvement

Introduction

In today's dynamic and competitive corporate environment, organizational success is contingent on the continuous development of human resources and the maximization of human performance (Mondy, Noe, & Premeaux, 2001). To facilitate these outcomes, more than 90% of Fortune 1000 companies utilize performance appraisal (Brutus, Petosa, & Aucoin, 2005). Performance appraisal is a formal process through which employees are evaluated by a judge, typically a supervisor, who assesses performance against a set of dimensions, assigns a score to performance, and communicates these results to the employee (DeNisi & Murphy, 2017). Performance appraisal has been of critical interest to academics and practitioners for more than a hundred years and the effectiveness of performance appraisal systems remains a vital topic in the field of Industrial Organizational Psychology (DeNisi & Murphy, 2017; Iqbal, Akbar, & Budhwar, 2015).

Despite a long term focus on performance appraisal, Adler et al. (2016) report that dissatisfaction with performance appraisal is at an all time high. Human resource managers feel that performance evaluation can be cumbersome, time consuming, and ultimately demonstrate few direct improvements on employee engagement or performance (Dorsey & Mueller-Hanson, 2017). These concerns have

led many companies to consider changes to their performance management systems, with a critical focus on the common practice of using ratings to evaluate employee performance (Adler et al., 2016). Performance ratings were developed due to the need to measure performance when objective measures were unavailable or deficient (Newman, Kinney, & Farr, 2004). Today, most organizations measure performance subjectively via ratings and performance appraisal typically involves rating an employee's performance on one or more dimensions using a 3-, 5-, or 7-point scale (Dorsey & Mueller-Hanson, 2017; Wall et al., 2004). Although common, this practice is a source of frustration for academics and practitioners as it has been studied and used for decades, often without favorable results (Bleckman et al., 2016).

For many organizations, performance appraisal is intended to be an instrument for performance improvement, which if used correctly, should motivate employees, result in positive behavior changes, and ultimately enhanced performance (Adler et al., 2016; Jansen & Vloeberghs, 1999). Despite this intention, the effectiveness of appraisal as a catalyst for performance improvement may be based on faulty assumptions and has long been debated. One key issue fueling this debate centers on the use of ratings to evaluate performance. This practice is often based on the assumption that

there is a relationship between the performance of employees and the ratings they receive and that rating feedback will have an effect on subsequent performance. In other words, that low performing employees will receive low ratings and, as a result, use this feedback to objectively improve performance (Adler et al., 2016; Dorsey & Mueller-Hanson, 2017). Yet, there is great controversy as to whether this assumption is warranted. Many executives do not believe that performance ratings are accurate reflections of performance (Dorsey & Mueller-Hanson, 2017) and, even if accurate, there is insufficient reason to believe that these ratings would be accepted and acted upon by the employee (Adler et al., 2016).

Consistent with these opinions, there is a lack of empirical evidence to support the assumption that rating feedback will translate into improvement in objective performance outcomes (Atwater, Waldman, & Brett, 2002). The few studies focused on this assumption have shown limited or no support that rating feedback impacts performance outcomes (e.g., Bernardin, Hagan, Ross, & Kane, 1995). Given that organizations typically provide employees with rating feedback with the intention of improving performance, it would seem that research investigating the effect of subjective rating feedback on objective performance outcomes would have the highest utility for practitioners. Contrary to this, two categories of research dominate

the literature. The first focuses on the effects of objective feedback on improvements in objective performance outcomes (see Davis, Carson, Ammeter, & Treadway, 2005; Mesch, Farh, & Podsakoff, 1994; Thorndike, 1927; Trowbridge & Cason, 1932). This traditional, predominantly laboratory-based approach defines feedback as knowledge of results: the type, extent, or direction of errors (Becker, 1978). While this type of experimental research yields many fruitful insights into the effectiveness of feedback interventions, its external validity is limited by the widespread use of subjective ratings as the basis for performance appraisal.

Thus, many researchers have shifted focus to field-based studies investigating the effects of subjective rating feedback on subjective performance improvements measured as changes in performance ratings over time (see Atwater, Roush, & Fischthal, 1995; Heslin & Latham, 2004; Smither et al., 1995; Reilly, Smither, & Vasilopoulos, 1996; Walker & Smither, 1999). On average, these studies have found a small, albeit significant positive effect on subsequent ratings after feedback (Byham & Weaver, 2005). While these studies provide evidence for feedback effectiveness, some researchers warn that comparing an individual's performance ratings before and after feedback is a deceptively simple approach to measuring performance improvement (Smither & Walker, 2001) and highlight that the lack of

focus on objective performance outcomes reduces the opportunity to illustrate a link between rating feedback and performance (Atwater, Waldman, & Brett, 2002; Church, 2000). Many authors have called for additional research that investigates the effect of rating feedback on objective measures of individual and organizational performance outcomes (Atwater et al., 2002; Atwater & Brett, 2005; Dominick, Reilly, & McGourty, 1997; Healy & Rose, 2003; Rotundo, 2002; Seifert, Yulk, & McDonald, 2003).

An additional concern regarding research common to the performance appraisal literature is that the frequent use of quasi-experimental designs, particularly the lack of random assignment to feedback conditions and no-feedback control groups, limit the interpretability of research findings (Atwater et al., 1995; Johnson & Ferstl, 1999; Smither, London, & Reilly, 2005; Smither & Walker, 2001; Walker & Smither, 1999). Due to organizational constraints, much research has defined feedback interventions as molar packages which consist of many different parts. Therefore, even if a positive effect is identified, it is often unclear whether this improvement can be attributed to feedback alone or another aspect of the intervention such as an awareness of appraisal criteria or development planning.

One aspect of the performance appraisal process that may contribute to performance improvement, independent of feedback, is knowledge of behavioral (appraisal) criteria - the criteria on which employees are evaluated. Simply participating in the appraisal process may familiarize employees with relevant behavioral criteria / expectations and help them focus on improvement even without feedback (Dalessio, 1998). Numerous researchers have found supportive evidence of this phenomenon, showing improvement in performance ratings as a result of familiarization with the appraisal instrument alone (Dominick et al., 1997; Smither et al., 1995; Van Velsor & Leslie, 1991). A second, commonly accepted, practice to supplement feedback is development planning. While familiarity with the appraisal instrument or feedback alone may be effective at increasing self-awareness, intentional change cannot be expected without commitment and planning. The importance of setting developmental goals is supported by Kluger and DeNisi's (1996) meta-analytic findings and numerous researchers have found that a combination of feedback and developmental goal setting had a larger effect on performance improvement than feedback alone (Hazucha, Hezler, & Schneider, 1993; Walker & Smither, 1999).

The present study was designed to replicate past research and to address gaps in the literature, using an experimental design to

directly test the assumption that rating feedback can lead to improvements in objective performance. Specifically, this research investigated the independent and combined effects of knowledge of behavioral criteria, rating feedback, and development planning on changes in subsequent performance ratings and objective performance. Based on previous research it was expected:

H1. There is a main effect of condition on posttest performance ratings, when pretest ratings are held constant such that (a) participants receiving knowledge of behavioral criteria obtain higher ratings than the control group; (b) participants receiving feedback obtain higher ratings than those receiving only knowledge of criteria; and (c) the addition of development planning leads to higher ratings compared to either knowledge of criteria or feedback alone.

H2. There is a main effect of condition on posttest group performance scores, when pretest scores are held constant such that (a) participants receiving knowledge of behavioral criteria achieve higher scores than the control group; (b) participants receiving feedback achieve higher scores than those receiving only knowledge of criteria; and (c) the addition of development planning leads to higher scores compared to either knowledge of criteria or feedback alone.

METHOD

PARTICIPANTS

Participants were 75 undergraduate students enrolled in Introduction to Psychology at a Northeastern University. Sixty one percent of participants were female. To volunteer, students selected one time slot which included the dates and times of two sessions. Based on their selection, participants were placed in three-person groups. Each of the 25 groups was then randomly assigned to one of the five conditions: (1) knowledge of behavioral criteria, (2) knowledge of behavioral criteria and development planning, (3) rating feedback, (4) rating feedback and development planning, or (5) no-treatment control group

Materials and Measures

Tasks. Participants completed two group consensus seeking tasks: Adventures in the Amazon (Ukens, 1998) and Lost in the Cradle of Gold (Ukens, 2005). A pilot study confirmed that the two tasks are equivalent in difficulty level. Both tasks required participants to read a scenario depicting an emergency situation and individually rank 15 items based on importance for survival. Upon completion, participants engaged in a group discussion to come to consensus regarding the prioritization of items. Group performance was calculated as the summed difference score of the group's prioritization of each item

compared to an optimal solution. Possible accuracy scores ranged from 0 (perfect) to 112. Tasks were chosen so that: (1) participants could be observed engaging in a group discussion, allowing for performance ratings and (2) both tasks have an optimal solution which enables the calculation of an objective performance score.

Rating form. The completion of each task required significant interaction and discussion among participants and therefore it was possible to observe and rate their behaviors while performing the task. Each group discussion was videotaped and later reviewed by trained assessors. The rating form used to subjectively evaluate participants' performance in the consensus seeking discussion was adapted from the Team Developer assessment program (McGourty & DeMeuse, 2001). To reduce the full list of 50 behavioral items, three I/O Psychology Master's students who served as raters in the experiment viewed a pilot study of participants completing the two tasks. The quality of each of the potential behavioral items was evaluated using five criteria proposed by Lovler, Rose, and Wesley (2002) (Appendix A). Twelve behavioral items, across four dimensions of team performance (collaboration, communication, decision making, and self-management), met all five criteria and were included in the final rating form (Appendix B). Participants in the experiment were

evaluated using a quality scale ranging from 1 - Inadequate to 7 - Excellent.

Intervention. Participants in the experimental conditions received variations of a comprehensive feedback and development planning guide adapted from the Team Developer assessment program. The complete guide consisted of a list of the 12 behavioral items included in the rating form, a feedback report including the participant's ratings on each behavioral item, developmental suggestions to improve each behavior, and a blank development planning worksheet which allowed participants to select three behaviors for improvement and choose two developmental suggestions for each behavior. Participants received variations of this guide, based on their condition, via email after the first experimental session. All participants were required to respond to the email to ensure treatment receipt.

Design and Procedure

The present study used a randomized pre-test post-test experimental design with a no-treatment control group. Each three-person group was randomly assigned to one of five intervention conditions. Groups participated in two group consensus seeking tasks, scheduled two weeks apart. The intervention was manipulated as an email, with variations of the feedback and development planning

guide, sent to participants two days after the first experimental session. Two dependent measures were collected after each experimental session, subjective individual performance ratings and objective group performance score. This allowed for the assessment of both subjective and objective performance improvements in each group.

Session one. In session one of the experiment, each group performed one of the consensus seeking tasks, which were counterbalanced to reduce order effects. Each session was videotaped. The participants first read the emergency scenario and performed the task individually. Upon completion, group members engaged in discussion to reach consensus on the prioritization of the 15 items. The accuracy of their prioritization was calculated and served as an objective measure of the group's performance at time 1 (pre-intervention). Following session one, trained raters observed videotapes of participants engaging in the group consensus seeking task. Each rater was randomly assigned one of the three participants to observe and evaluated that participant on 12 behavioral criteria. These ratings were the basis of the feedback intervention. The average rating for each participant was calculated and served as the subjective measure of performance at time 1 (pre-intervention).

Intervention. Two days after session one of the experiment, participants were emailed variations of the feedback and development planning guide based on their experimental condition as detailed:

Design and Procedure

The present study used a randomized pre-test post-test experimental design with a no-treatment control group. Each three-person group was randomly assigned to one of five intervention conditions. Groups participated in two group consensus seeking tasks, scheduled two weeks apart. The intervention was manipulated as an email, with variations of the feedback and development planning guide, sent to participants two days after the first experimental session. Two dependent measures were collected after each experimental session, subjective individual performance ratings and objective group performance score. This allowed for the assessment of both subjective and objective performance improvements in each group.

Session one. In session one of the experiment, each group performed one of the consensus seeking tasks, which were counterbalanced to reduce order effects. Each session was videotaped. The participants first read the emergency scenario and performed the task individually. Upon completion, group members engaged in discussion to reach consensus on the prioritization of the 15 items. The accuracy of their prioritization was calculated and served as an objective measure of the group's performance at time 1 (pre-

intervention). Following session one, trained raters observed videotapes of participants engaging in the group consensus seeking task. Each rater was randomly assigned one of the three participants to observe and evaluated that participant on 12 behavioral criteria. These ratings were the basis of the feedback intervention. The average rating for each participant was calculated and served as the subjective measure of performance at time 1 (pre-intervention).

Intervention. Two days after session one of the experiment, participants were emailed variations of the feedback and development planning guide based on their experimental condition as detailed:

Knowledge of behavioral criteria. The email included a list of 12 behaviors from the rating form. Participants were told that these behaviors were essential to effective group performance. Although they did not receive feedback on their performance, they were instructed to review the list before the next experimental session.

Knowledge of behavioral criteria and development planning. In addition to knowledge of behavioral criteria, these participants received a development planning guide and were instructed to review the list of behaviors, think about their performance in the task, and create a development plan. The guide included developmental suggestions and a blank development planning worksheet which allowed participants to target three behaviors for improvement by selecting two developmental activities to help improve

each behavior. Participants were instructed to use the developmental suggestions to improve their behavior before Session 2 of the experiment.

Rating feedback. Participants were informed that their performance in the group consensus seeking task had been rated by trained assessors on 12 behaviors essential to team performance. The email included a feedback report detailing their ratings on each of the 12 behaviors and instructions on how to interpret the report.

Rating feedback and development planning. In addition to feedback, participants received a development planning guide and were told to review their feedback report and select the three lowest-rated behaviors for improvement. The guide included developmental suggestions and a blank development planning worksheet which allowed participants to select two developmental activities to help improve each of the three lowest-rated behaviors. They were instructed to work on these behaviors before the next experimental session.

No-treatment control. Participants received a reminder email about Session 2 with no additional instructions.

Session two. Two weeks following Session 1, the three-person groups returned to complete the second consensus seeking task. This time period was selected to allow participants in the development planning conditions time to work on improving their

behavior. Session 2 followed the same procedure as Session 1. The accuracy of the group's prioritization on the second task was calculated and served as an objective measure of the group's performance at time 2 (post-intervention). Following Session 2, trained assessors reviewed the videotaped experimental sessions and again were randomly assigned to rate one participant. The average behavioral rating for each participant was calculated and served as a subjective measure of performance at time 2 (post-intervention). No additional feedback was provided.

Results

Hypothesis 1: Subjective Performance Ratings

To analyze improvement in subjective ratings, an overall rating was calculated for each participant across all 12 behaviors at Time 1 and Time 2. Before aggregating items, the inter-item reliability for the 12 behavioral items was computed ($\alpha = .78$). Descriptive statistics for ratings, including improvement from Time 1 to Time 2, are shown in Table 1. Mean improvement scores calculated as the average rating at Time 2 - Time 1 indicate that participants in the control condition were rated lower at Time 2 ($M = -0.28$) while those receiving a combination of feedback and development planning showed the greatest improvement ($M = 0.54$). Due to the relatively small sample

size, ANCOVA was chosen to assess differences in behavioral ratings at the posttest, while holding pretest differences constant. Results of previous studies suggest that in randomized trials, ANCOVA often provides a more powerful analysis compared to similar statistical techniques including gain score analysis (Oakes & Feldman, 2001).

Table 1
Results of Descriptive Statistics for Subjective Performance Ratings

Condition	n	Overall Rating					
		Time 1		Time 2		Improvement	
		M	SD	M	SD	M	SD
Condition 1 <i>Knowledge of Behavioral Criteria</i>	15	3.72	0.54	3.62	0.49	-0.10	0.60
Condition 2 <i>Knowledge of Criteria & Development Planning</i>	15	3.82	0.64	3.87	0.64	0.04	0.73
Condition 3 <i>Feedback</i>	15	3.66	0.47	4.08	0.68	0.42	0.66
Condition 4 <i>Feedback & Development Planning</i>	15	3.79	0.64	4.33	0.84	0.54	0.59
Condition 5 <i>Control</i>	15	3.63	0.54	3.34	0.57	-0.28	0.60

Note: Overall rating is an average of participant's ratings on all 12 behavioral items. The possible range of overall rating was 1 (inadequate) to 7 (excellent). A positive improvement indicates higher ratings at Time 2.

Consistent with Hypothesis 1, results of ANCOVA indicated a significant difference among posttest ratings across conditions, $F(4, 69) = 5.77, p = 0.00$ (Table 2). Results of Fisher's Least Significant Difference (LSD) pairwise comparisons provided a more detailed examination of the differences in posttest ratings. Contrary to Hypothesis 1a, participants who received only knowledge of criteria

were not rated higher than participants in the control condition. Providing support for Hypothesis 1b, participants who received behavioral feedback did obtain significantly higher posttest ratings compared to those in the control condition and who received only knowledge of criteria. Finally, the addition of development planning did lead to slightly higher mean posttest ratings than either knowledge of behaviors or feedback alone. However, in contrast to Hypothesis 1c, these differences were not significant. The addition of development planning did not result in significant improvements in behavioral ratings compared to knowledge of criteria or feedback alone.

Table 2
Results of ANCOVA for Subjective Performance Ratings

	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>Effect Size</i>	<i>Observed Power</i>
Condition	4	1.96	5.77	0.00	0.25	0.98
Error	69	0.34				

Hypothesis 2: Objective Performance Scores

Descriptive statistics for objective group performance scores, including improvement from Time 1 to Time 2, are shown in Table 3. Note that while average subjective ratings were calculated for each individual participant, only one objective performance score could be calculated per group. Lower scores indicate greater accuracy, thus

negative improvement scores demonstrate better performance at Time 2. An examination of mean improvement scores reveals that the control condition performed worse at Time 2 ($M = 8.80$), while participants receiving only behavioral feedback showed the greatest improvement in objective performance ($M = -6.80$).

Table 3
Results of Descriptive Statistics for Group Performance Scores

Condition	n	Group Performance Score					
		Time 1		Time 2		Improvement	
		M	SD	M	SD	M	SD
Condition 1 <i>Knowledge of Behavioral Criteria</i>	5	58.00	13.19	62.40	4.34	4.40	15.65
Condition 2 <i>Knowledge of Criteria & Development Planning</i>	5	56.00	11.22	53.20	1.79	-2.80	12.13
Condition 3 <i>Feedback</i>	5	59.60	9.53	52.80	3.35	-6.80	9.58
Condition 4 <i>Feedback & Development Planning</i>	5	60.40	2.61	54.40	4.98	-6.00	6.78
Condition 5 <i>Control</i>	5	53.60	6.69	62.40	6.84	8.80	6.10

Note: The possible range of group performance score was 0 (perfect score) 112. A negative improvement indicates greater accuracy at Time 2.

ANCOVA was again performed to assess the differences in group performance score at Time 2, with pretest scores held constant. Consistent with Hypothesis 2, results indicated a significant difference between posttest scores across conditions, $F(4, 19) = 5.31, p = 0.01$ (Table 4). LSD pairwise comparisons reveal a pattern of results similar to improvements in behavioral ratings. As in the previous analysis, Hypothesis 2a was not supported. Participants who received

knowledge of criteria were not significantly more accurate than the control condition. Consistent with Hypothesis 2b, participants who received behavioral feedback achieved significantly greater posttest performance scores than the control condition or those receiving knowledge of criteria. Providing partial support for Hypothesis 2c, the effects of development planning were mixed. As predicted, those receiving a combination of knowledge of criteria and development planning outperformed those receiving only knowledge of criteria. However, similar to the results for behavioral ratings, a combination of behavioral feedback and development planning did not lead to greater accuracy than feedback alone.

Table 4
Results of ANCOVA for Group Performance Scores

	<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	<i>Effect Size</i>	<i>Observed Power</i>
Condition	4	116.83	5.31	0.01	0.53	0.92
Error	19	21.99				

Discussion

Organizations continue to critically focus on the effectiveness of their performance appraisal systems, particularly the use of performance ratings as the foundation of these systems. The long-standing dissatisfaction with the use of ratings and lack of empirical evidence supporting the assumption that rating feedback will impact

performance has left many organizations considering the elimination of performance ratings and, in some cases, annual appraisals altogether (Adler et al., 2016; Cappelli & Tavis, 2016). Estimates, however, suggest that most organizations are enacting less extreme changes to their performance systems. Many researchers and practitioners acknowledge a need for the use of performance ratings and believe that eliminating such systems is not the panacea (Adler et al., 2016; Cappelli & Tavis, 2016; Gorman, Cunningham, Bergman, & Meriac, 2016). At this critical time in the history of performance appraisal, research is needed to test the assumptions on which traditional performance appraisal systems were developed. The present experiment adds to the existing literature by directly examining the impact of rating feedback on both subjective and objective performance improvements.

Explanation of Results

The first goal was to replicate results of past research by measuring changes in subjective behavioral ratings following the intervention. Feedback was found to be a necessary and sufficient condition for improvement in behavioral ratings. Contrary to expectations, participants who received only knowledge of behavioral

criteria did not show greater improvement than the no-treatment control condition. Participants who received rating feedback outperformed those in the control condition and those receiving only knowledge of criteria. This result is consistent with hypotheses, as well as findings of prior research, touting the positive impact of feedback on improvements in behavioral ratings. Further, feedback was found to be a sufficient condition for improvement as the addition of development planning did not have a significant effect on improvements in behavioral ratings.

Second, this experiment extended results of past research by directly measuring the impact of behavioral rating feedback on objective performance improvement. Results for improvement in objective performance scores closely mirrored those for changes in subjective performance ratings following the intervention. Specifically, results revealed that knowledge of behavioral criteria was insufficient to cause an improvement in objective performance. Feedback, again, was found to be a necessary condition for improvement and participants in the feedback condition achieved the greatest improvements in objective performance scores following the intervention. Results for development planning were mixed. While the addition of development planning significantly improved post-test

scores for the knowledge of behavioral criteria condition, it did not have a significant impact for participants receiving feedback.

Limitations

Contrary to expectations, knowledge of behavioral criteria and development planning were not found to have a significant impact on performance improvement. This finding is inconsistent with the results of prior research. These discrepancies may be due to methodological differences and limitations in the present study. One primary difference is the use of a laboratory-based intervention as compared to the field studies commonly used to evaluate performance appraisal initiatives. Much prior research investigating the effectiveness of feedback interventions has been conducted in a field setting where the implementation of performance appraisal suggests that the organization is making an effort to foster development. Because such development may lead to a number of desired consequences such as recognition, bonuses, and promotion, participants are likely more motivated to improve as compared to laboratory-based interventions, even in the absence of direct feedback.

Several other methodological characteristics of the present study may help to explain differences in results. Past research has shown that individuals who simply participate in a performance appraisal process achieve similar improvements in behavioral ratings as those who actually receive feedback (Dominick et al., 1997; Reilly et al., 1996; Smither et al., 1995). To explain this finding, researchers suggest that familiarization with relevant behavioral criteria may be sufficient to cause spontaneous goal setting even without feedback (Locke & Latham, 1990; Delassio, 1998). In much of this prior research, participants were not only exposed to behavioral criteria but were involved in the appraisal process, using those criteria to rate their manager or peers. Hence, mere exposure to the behavioral criteria may not be sufficient for performance improvement. Participation in an appraisal initiative, such as through the provision of ratings, may be necessary to yield improvements in the absence of feedback.

The contradictory findings regarding development planning may also be partially explained by a methodological limitation. Participants engaged in individual development planning with the purpose of improving group performance. However, research has suggested that individual goal setting leads to feeling less like a team, less cooperative, and more competitive toward group members. Further, individuals have reported forming significantly less cooperative strategies and more competitive strategies in which they tried to

outperform their group members which can lead to decreased group performance (Mitchell & Silver, 1990). Thus setting individual goals for an interdependent task may have masked the effects of development planning in the present experiment.

Conclusions

Overall, rating feedback was found to have the greatest impact on both improvements in subsequent behavioral ratings and objective performance scores. Most notably, this experiment provides support for the assumption that rating feedback can impact objective performance. Participants who received behavioral rating feedback obtained significantly higher objective performance scores on a group consensus seeking task than those who did not receive feedback. At a time when many organizations are considering significant changes to their performance appraisal systems, it is critical that researchers and practitioners continue to examine the effectiveness of rating feedback on performance improvement. Future field research is needed to generalize results of the present study in an organizational setting.

References

Adler, S., Campion M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E.D. (2016). Getting rid of performance ratings: Genius or folly? A debate. *Industrial / Organizational Psychology*, 9(2), 219-252. doi: 10.1017/iop.2015.106

- Atwater, L.E., & Brett, J.F. (2005). Antecedents and consequences of reactions to developmental 360 feedback. *Journal of Vocational Behavior*, 66, 532-548. doi: 10.1016/j.jvb.2004.05.003
- Atwater, L.A., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self and follower ratings of leadership. *Personnel Psychology*, 48, 35-60. doi: 10.1111/j.1744-6570.1995.tb01745.x
- Atwater, L.E., Waldman, D.A., & Brett, J.F. (2002). Understanding and optimizing multisource feedback. *Human Resource Management*, 41(2), 193-208. doi: 10.1002/hrm.10031
- Becker, L.J. (1978). Joint effect of feedback and goal setting on performance: A field study of residential energy conservation. *Journal of Applied Psychology*, 63(4), 428-433. doi: 10.1037/0021-9010.63.4.428
- Bernardin, H.J., Hagan, C., Ross, S., & Kane, J.S. (1995, May). The effects of a 360 degree appraisal system on managerial performance. In W.W. Tornow (Chair), *Upward feedback: The ups and downs of it*. Symposium conducted at the 10th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
- Bleckman, A.M., Guarino, S.N., Russell, W., Toomey, E.C., Werth, P.M., Whitaker, V.L., & Rudolph, C.W. (2016). The future of performance ratings: Collected thoughts from six emerging scholars. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 9(2), 334-337. doi: 10.1017/iop.2016.20
- Brutus, S., Petosa, S., & Aucoin, E. (2005). Who will evaluate me? Rater selection in multi source assessment contexts. *International Journal of Selection and Assessment*, 13(2), 129-138. doi: 10.1111/j.0965-075X.2005.00307.x
- Byham, W.C., & Weaver, P.S. (2005). *Multisource feedback that effects changes in leaders' behavior*. Pittsburgh, PA: Development Dimensions International, Inc. Retrieved from: <http://www.performance-vision.com/step/webinars/Multi-RaterROI/DDI-MultiRaterFeedbackOptions.pdf>
- Capelli, P., & Tavis, A. (2016, October). The performance management revolution. *Harvard Business Review*, 94(10), 58-67. Retrieved from: <https://hbr.org/2016/10/the-performance-management-revolution#>
- Church, A. (2000). Do higher performing managers actually receive better ratings? A validation of multi-rater assessment methodology. *Consulting Psychology Journal: Practice and Research*, 52, 99-116. doi: 10.1037/1061-4087.52.2.99
- Dalessio, A.T. (1998). Using multisource feedback for employee development and personnel decisions. In Smither, J.W. (ed.) *Performance Appraisal: State of the Art in Practice*. Jossey-Bass: San Francisco, CA.

- Davis, W.D., Carson, C.M., Ammeter, A.P., & Treadway, D.C. (2005). The interactive effects of goal orientation and feedback specificity on task performance. *Human Performance, 18*(4), 409-426. doi: 10.1207/s15327043hup1804_7
- DeNisi, A.S. & Murphy, K.R. (2017). Performance appraisal and performance management: 100 years of progress? *Journal of Applied Psychology, 102*(3), 420-433. doi: 0.1037/apl0000085
- Dominick, P.G., Reilly, R.R., & McGourty, J.W. (1997). The effects of peer feedback on team member behavior. *Group and Organization Management, 22*, 508-520. doi: 10.1177/1059601197224006
- Gorman, C.A., Cunningham, C., Bergman, S.M., & Meriac, J.P. (2016). Time to change the bathwater: Correcting misconceptions about performance ratings. *Industrial and Organizational Psychology, 9*(3), 314-322. doi: 10.1017/iop.2016.17
- Hazucha, J.F., Hezlett, S.A., & Schneider, R.J. (1993). The impact of 360-degree feedback on management skills development. *Human Resource Management, 32*, 325-351. doi: 10.1002/hrm.3930320210
- Healy, M.C., & Rose, D.S. (2003). *Validation of a 360 degree feedback instrument against sales: Content matters*. Poster presented at the 18th Annual Conference for the Society of Industrial / Organizational Psychology, Orlando, FL.
- Heslin, P.A. & Latham, G.P. (2004). The effect of upward feedback on managerial behavior. *Applied Psychology: An International Review, 53*(1), 23-37. doi: 10.1111/j.1464-0597.2004.00159.x
- Iqbal, M.Z., Akbar, S., & Budhwar, P. (2015). Effectiveness of performance appraisal: An integrated framework. *International Journal of Management Reviews, 17*, 510-533. doi: 10.1111/ijmr.12050
- Jansen, P. & Vloeberghs, D. (1999). Multirater feedback methods: Personal and organizational implications. *Journal of Managerial Psychology, 14*, 455-476. doi: 10.1108/02683949910287903
- Johnson, J.W., & Ferstl, K.L. (1999). The effects of interrater and self-other agreement on performance improvement following upward feedback. *Personnel Psychology, 52*, 271-303. doi: 10.1111/j.1744-6570.1999.tb00162.x
- Kluger, A.N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284. doi: 10.1037/0033-2909.119.2.254

- Lovler, R. Rose, M. & Wesley, S. (2002, April). *Assessment centers as tests: Implications for analyzing assessment center contrast-related validity*. Poster presented at the 17th Annual Conference for the Society for Industrial and Organizational Psychology, Toronto, Canada.
- McGourty, J., & DeMeuse, K.P. (2001). *The team developer: An assessment and skill building program student guidebook*. New York, NY: John Wiley & Sons.
- Mesch, D.J., Farh, J., & Podsakoff, P.M. (1994). Effects of feedback sign on group goal setting, strategies, and performance. *Group and Organizational Management*, 19(3), 309-333. doi: [10.1177/1059601194193006](https://doi.org/10.1177/1059601194193006)
- Mitchell, T.R. & Silver, W.S. (1990). Individual and group goals when workers are interdependent: Effects on task strategies and performance. *Journal of Applied Psychology*, 75(2), 185-193.
- Mondy, R.W., Noe, R.M., & Premeaux, S.R. (2001). *Human Resource Management*. Upper Saddle River, NJ: Prentice Hall.
- Newman, D.A., Kinney, T., & Farr, J.L. (2004). Job performance ratings. In J.C. Thomas (Ed.), *Comprehensive Handbook of Psychological Assessment, 4*, 373-388. New York, NY: Wiley & Sons.
- Oakes, J. M., & Feldman, H. A. (2001). Statistical power for non-equivalent pretest posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review*, 25, 3-28. doi: [10.1177/0193841X0102500101](https://doi.org/10.1177/0193841X0102500101)
- Reilly, R.R., Smither, J.W., & Vasilopoulos, N.L. (1996). A longitudinal study of upward feedback. *Personnel Psychology*, 49, 599-612. doi: [10.1111/j.1744-6570.1996.tb01586.x](https://doi.org/10.1111/j.1744-6570.1996.tb01586.x)
- Rotundo, M. (2002). *Defining and measuring individual level job performance: A review and integration*. Retrieved from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.7597&rep=rep1&type=pdf>
- Seifert, C.F., Yukl, G., & McDonald, R.A. (2003). Effects of multisource feedback and a feedback facilitator on the influence behavior of managers toward subordinates. *Journal of Applied Psychology*, 88(3), 561-569. doi: [10.1037/0021-9010.88.3.561](https://doi.org/10.1037/0021-9010.88.3.561)
- Smither, J.W., London, M., & Reilly, R.R. (2005). Does performance improve following multi-source feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58, 33-66. doi: [10.1111/j.1744-6570.2005.514_1.x](https://doi.org/10.1111/j.1744-6570.2005.514_1.x)
- Smither, J., London, M., Vasilopoulos, N., Reilly, R., Millsap, R., & Salvemini, N. (1995). An examination of the effects of an upward feedback program over time. *Personnel Psychology*, 48, 1-34. doi: [10.1111/j.1744-6570.1995.tb01744.x](https://doi.org/10.1111/j.1744-6570.1995.tb01744.x)
- Smither, J.W., & Walker, A.G. (2001). Measuring the impact of multi-source feedback. In D.W. Bracken, C.W. Timmreck, & A.H. Church (Eds.),

- The Handbook of Multisource Feedback*. San Francisco, CA: Jossey-Bass.
- Thorndike, E.L. (1927). The law of effect. *American Journal of Psychology*, 39, 212-222. doi: 10.2307/1415413
- Trowbridge, M.A., & Cason, H. (1932). An experimental study of Thorndike's theory of learning. *Journal of General Psychology*, 7, 245-260. doi: [10.1080/00221309.1932.9918465](https://doi.org/10.1080/00221309.1932.9918465)
- Ukens, L.L. (1998). *Adventures in the amazon*. San Francisco, CA: Pfeiffer.
- Ukens, L.L. (2004). *Lost in the cradle of gold*. San Francisco, CA: Pfeiffer.
- Van Velsor, E., & Leslie, J.B. (1991). *Feedback to managers, volume II: A review and comparison of 16 multirater feedback instruments*. Greensboro, NC: Center for Creative Leadership.
- Walker, A.G., & Smither, J.W. (1999). A five year study of upward feedback: What managers do with their results matters. *Personnel Psychology*, 52, 393-423. doi: 10.1111/j.1744-6570.1999.tb00166.x
- Wall, T.D., Michie, J., Patterson, M., Wood, S.J., Sheehan, M., Clegg, C.W., & West, M. (2004). On the validity of subjective measures of company performance. *Personnel Psychology*, 57, 95-118. doi: 10.1111/j.1744-6570.2004.tb02485.x

Appendix A
Criteria Proposed by Lovler, et al. (2002)

1	The meaning of the behavioral statement was clear.
2	The assessor could be expected to reliability distinguish between multiple levels of performance on the behavior.
3	Candidates could be expected to vary in their performance on the behavior.
4	There was a clear opportunity to observe the behavior in the exercise.

5	The behavioral statement measured a single behavior.
---	--

Appendix B Final Rating Form

Based on your observation of the participant's behavior in the group consensus seeking task, evaluate his/her level of performance on each behavior using the 7 point rating scale below.

I. Collaboration

Demonstrating a commitment to the team's overall purpose, helping team members to identify mutual objectives, working cooperatively and constructively with others in the team, actively participating in team activities, showing support and encouragement for fellow team members.

	1	2	3	4	5	6	7
	Inadequate	Poor	Fair	Average	Good	Very Good	Excellent
Acknowledges issues that the team needs to confront & resolve....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Encourages participation among all participants.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shares information with others.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reinforces the contributions of others.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

II. Communication

Helping to sustain an environment where people feel free to speak candidly, articulating ideas clearly and concisely, listening and demonstrating an understanding of others' perspectives.

	1	2	3	4	5	6	7
	Inadequate	Poor	Fair	Average	Good	Very Good	Excellent
Restates what has been said to show understanding.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uses facts to get points across.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

III. Decision Making

Gathering information and weighing alternatives when addressing an issue, working with the team toward resolution, promoting innovative thinking, ensuring that a rationale forms the basis for the decision made

	1	2	3	4	5	6	7
	Inadequate	Poor	Fair	Average	Good	Very Good	Excellent
Generates new ideas.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Discourages others from rushing to conclusions without facts.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

IV. Self Management

Utilizing appropriate styles, methods and procedures to direct individuals and the team toward goal achievement, modeling and modifying behavior as required to achieve results while being sensitive to individual and group processes.

	1	2	3	4	5	6	7
	Inadequate	Poor	Fair	Average	Good	Very Good	Excellent
Puts top priority on getting results.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Stays focused on the task during work sessions.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Suggests ways to proceed during work sessions.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reviews progress through work sessions.....	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>