THE EFFECT OF MULTI-RATER CONSENSUS ON PERFORMANCE RATING ACCURACY

Carrie Ann Picardi*

**Assistant Professor of Management, Ernest C. Trefz School of Business, University of Bridgeport, United States of America. Email: cpicardi@bridgeport.edu*

Abstract *This study examined the extent that consensus affects performance rating accuracy. Participants (n=96) viewed a video depicting teams working on a problem-solving exercise. The ratees were evaluated on behaviours within three performance dimensions: verbal communication, collaboration, and decision making. Rating accuracy across three conditions (consensus, discussion without consensus, control) was calculated using Cronbach's (1955) accuracy indexes: elevation, differential elevation, stereotype accuracy, and differential accuracy. It was hypothesized that ratings provided by participants in the consensus condition would demonstrate the highest degree of accuracy. Findings in support of this hypothesis provided justification for use of consensus-driven multirater teams for performance evaluation. Moreover, organisations that incorporate a multi-rater strategy in the performance management process may consider a consensus approach rather than individual ratings submission to enhance accuracy.*

Keywords: *Employee, Performance, Evaluation, Appraisal, Rating, Accuracy*

Introduction

Performance management is acknowledged by many organisations to be a valuable and necessary component in the evaluation of employees, and impacts other facets of the employee lifecycle, including training and development, compensation, rewards and incentives, and disciplinary action. The domain of performance management is comprised of several distinct areas: (1) assessment, (2) feedback, and (3) reactions, with each area consisting of specific strategies and processes. A critical goal of organisations is to establish a relationship between employees' assessed behaviours and performance outcomes. An appropriately designed and implemented performance management system in an organisation, with adequate attention to key elements such as competency and rating scale development, alignment of performance outcomes to organisational objectives, rating accuracy, feedback, goal setting, and development opportunities, may offer substantial value in terms of a high-performing workforce and successful achievement of organisational objectives (London, Mone, & Scott, 2004).

Evaluation of Job-Related Behaviours

Measurement of job-specific behaviours and performance outcomes is typically measured in two ways: (1) objectively, by identifying the presence of the behaviour and determining its frequency, and (2) subjectively, by using one's judgment to examine the quality of the behaviour demonstrated. The examination of an objective measure requires a focus on *behavioural accuracy* and the examination of a subjective measure requires a focus on *rating accuracy*. Behavioural accuracy refers to the correct identification of whether a behaviour occurred, and rating accuracy refers to the appropriate rating of a behaviour and the extent to which it matches a standardised rating score. The ability to make accurate performance ratings is an integral part of the assessment process, the last step in a three-step process that is most commonly conducted by a job incumbent's supervisor: information collection, processing, and judgment/rating (Ilgen & Favero, 1985). Researchers (Borman, 1977; Murphy & Balzer, 1989; Sulsky & Balzer, 1990; Roch, 2006) generally concur that, while behavioural and rating accuracy share a common purpose, rating accuracy through subjective measurement techniques is important in the research of employee performance evaluation because rating accuracy has the potential for greater external validity and therefore organisational utility than behavioural accuracy. These researchers have studied subjective measurement by examining variables that may affect rating accuracy, which is the focus of the present study.

Rating Accuracy in Performance Decisions

The ability to accurately rate individual work performance across myriad responsibilities or duties is particularly critical when high stakes decisions, such as selection and promotion, are made based on the ratings. Regardless of the type of personnel decision, two of the most critical factors for adequate rating accuracy are (1) appropriate and relevant criteria with which to evaluate the ratee's performance, typically obtained through a comprehensive job analysis and examination of a job's key duties and responsibilities as well as incumbent goals, and (2) appropriate raters who understand the process and performance standards, and also are familiar with the ratee and have enough opportunity to observe performance (Barnes-Farrell & Lynch, 2003). Inclusion of these factors will ensure standardisation of the steps comprising the performance evaluation process, thus improving rating accuracy through both sufficient criterion validity and interrater reliability.

A substantial amount of performance management research (Tetlock, 1985; Salvemini, Reilly, & Smither, 1993; DeNisi & Peters, 1996, Roch, 2006) has involved the examination of rating accuracy across a variety of different types and number of ratees and/or raters, performance criteria, and contextual factors. Martell and Borg (1993) studied behavioural rating accuracy of groups of raters versus individual raters, and based the study on two independent components believed to be important in recognizing key performance indicators and work behaviours: memory sensitivity, the ability to remember behaviours, and decision criteria, the ability to correctly determine whether or not a behaviour occurred when there is any degree of uncertainty (Lord, 1985; Martell & Guzzo, 1991). The researchers found that rater groups demonstrated greater memory sensitivity than individuals in the delayed rating condition and that the rater groups used a more liberal decision criterion than individual raters, findings that may provide insight into determining the appropriateness of using rater groups in the evaluation of job performance. Roch (2006) also examined behavioural accuracy and rating accuracy, and the extent to which group discussion and consensus affect behavioural accuracy and rating accuracy of performance ratings. Roch's research findings demonstrated significantly higher memory strength in the discussion-only and consensus conditions than the control condition and significant improvements in both behavioural accuracy and rating accuracy after reaching consensus, which provided support for the idea that anticipation of group discussion can increase behavioural accuracy as well as for the idea that consensus can improve both behavioural accuracy and rating accuracy. It can be inferred from these findings that discussion of ratings until consensus is reached may be a superior method for improving behavioural and rating accuracy compared to discussion without consensus (Roch, 2006).

Research in the area of utilisation of rater groups for performance evaluation is critical to organisations because the number of raters participating in a performance evaluation process may vary from one individual to several, both internal and external to the organisation, depending on resources and contextual needs. The individual rater model, in which one rater evaluates a job incumbent's performance over a specified period of time, has traditionally been a common organisational practice. The rater is typically the incumbent's immediate supervisor, and the evaluation process itself is typically a formal performance appraisal activity driven by a time milestone (e.g., three months, one year). While the single rater approach is the accepted practice in many organisations, a growing number of organisations are implementing a multirater performance feedback scenario in which relevant individuals, including subordinates, peers, fellow team/committee members, and customers/clients, provide ratings on an individual's job performance across a variety of work contexts (Smither, London, & Reilly, 2005). While organisations using multirater, or multisource, feedback typically use the information for employee development and goal setting purposes, some organisations do use the information for selection, promotion, and other high stakes personnel decisions. Organisational use of performance evaluation information for high stakes personnel decisions further emphasizes the importance of overall rating accuracy on various performance dimensions obtained through myriad sources.

The focus of the present research is on the implications of multi-rater consensus on performance rating accuracy. The inclusion of multiple raters, such as peers, managers, and/or customers, who may interact with the ratee in a variety of workplace situations, will likely have different perspectives on the ratee's performance based on their unique observations and interactions with the rate in certain contexts. Roch (2006) examined behavioural and rating accuracy in the context of both a discussion-only condition and a discussion with consensus condition in addition to the control condition; similarly the same three conditions will be implemented in the present study which will attempt to provide support for Roch's (2006) findings pertaining to rating accuracy of consensus-driven groups. It is proposed that a consensus requirement for the group, along with appropriate rating accuracy measures, will offer a feasible method for improvement and enhanced value to the performance assessment process. The following hypothesis has been developed:

Hypothesis 1: Individuals in the consensus rating condition will have greater rating accuracy in their consensus ratings than individuals in the no consensus rating condition or the

control condition will have in their individual summary ratings.

Methods

Research Participants

Participants (n = 96) were graduate students of a liberal arts college in the Northeast U.S., enrolled in several different psychology Masters degree programmes.

Materials

A 10-minute video recording, developed by Niemen-Gonder (2006), of a 3-person group working together on a problem-solving simulation exercise was shown to participants during the rating sessions. The use of video rather than written vignettes is a method initiated in performance rating research by Borman (1978) that gained support in subsequent studies (Martell & Borg, 1993; Roch, 2006) as a reasonable improvement in the generalisability of rating accuracy findings to applied settings. The three individuals who had been videotaped while working together on the simulation exercise served as the ratees, whose behaviours demonstrated across the performance dimensions of collaboration, verbal communication, and decision making, would be evaluated by the participants. These dimensions of performance have been consistently examined in research examining team decision-making and effectiveness in a variety of contexts (Mitchell & Silver, 1990; McGourty, 2001; Nieman-Gonder, 2006).

Measures

Independent Variable

The independent variable was the rating consensus requirement, and there were three levels: a consensus condition, a discussion-only condition, and a control condition. Though the participants in the discussion-only and control conditions made their ratings independently, they were also placed in groups of three raters, in alignment with the same setting in which the consensus rating condition participants were placed.

Dependent Variables

We included four dependent variables, using Cronbach's (1955) four accuracy indexes, which are predominantly used by researchers examining performance rating accuracy (Borman, 1977; Murphy & Balzer, 1989; Smither, Barry, & Reilly, 1989; Roach & Gupta, 1992; Roch, 2006) and have received significant empirical support as direct measures of the accuracy of subjective performance ratings:

Elevation (E): The accuracy of the average rating across all ratees and performance dimensions.

Differential Elevation (DE): The accuracy of a rating given to a specific ratee across all job dimensions.

Stereotype Accuracy (SA): The accuracy of a rating given to a specific job dimension across all ratees.

Differential Accuracy (DA): The accuracy of a rating given to a specific ratee on a specific performance dimension.

We followed the methodology and formulas provided by Sulsky and Balzer (1988) to calculate each of these accuracy indexes. Each accuracy measure requires rating scores provided by designated raters (i.e., observed scores) and standardized rating scores provided by one or more trained expert raters (i.e., true scores). Guion (1965) proposed that true scores should be used as an acceptable standard for a specified performance dimension to determine the degree of accuracy of observed ratings provided by a rater. True scores are typically calculated as the mean rating score across all expert raters for a specific individual ratee on a particular performance dimension. The closer the observed ratings are to the true scores, the greater the degree of confidence in the accuracy of the observed ratings. True scores have been incorporated by researchers (Woehr, 1994, Roch, 2006) in their studies of performance rating accuracy, and are a critical component in this method of measurement because such a standardized set of rating scores for all ratees and performance dimensions is needed to determine the degree of accuracy of the observed scores obtained.

We derived true scores from trained subject matter expert raters, consistent with the methodology of other researchers (Borman, 1977; Roach & Gupta, 1992; Woehr, 1994; Roch, 2006) incorporating true scores in their respective studies. We used the frame-of-reference training method for developing ten doctoral candidates into expert raters, which involved training on the three specified performance dimensions and their corresponding behaviours, and how to rate the behaviours demonstrated along the seven-point rating scale (e.g., what level of the demonstrated behaviours constitutes "excellent" performance on a particular dimension). The ratings provided by the expert raters were computed into pooled average scores and the data were used as true score ratings for each ratee on each behavioural performance dimension.

Procedures

The participants were randomly assigned to groups of three in one of the three experimental conditions. The participants

in each session viewed a video that presented a group of three individuals (ratees) working through a problem-solving simulation exercise. The same video was shown to all participant groups in all three experimental conditions. While viewing the video, the participants provided rating scores for the level of performance of each of the three specified performance dimensions for each ratee in the video. All participants used a seven-point rating scale, with a rating of 1 for inadequate performance and a rating of 7 for excellent performance, for each of the three performance dimensions and each of the three ratees.

The experimenter provided instructions specific to each group's condition between each session's two rating periods. Therefore, the groups were not aware of the necessary actions pertinent to their respective condition until after the first set of individual ratings were completed. Participants in the consensus condition were required to discuss their individual ratings with their fellow group members and reach consensus on one collective second set of ratings. Participants in the discussion with no consensus condition were required to discuss their individual ratings with their fellow group members and, after several minutes of discussion-free individual reflection, independently make a second set of summary rating scores. Participants in the control condition were required to complete a 10-item questionnaire consisting of opinion questions regarding the video presentation, and then independently make a second set of summary rating scores.

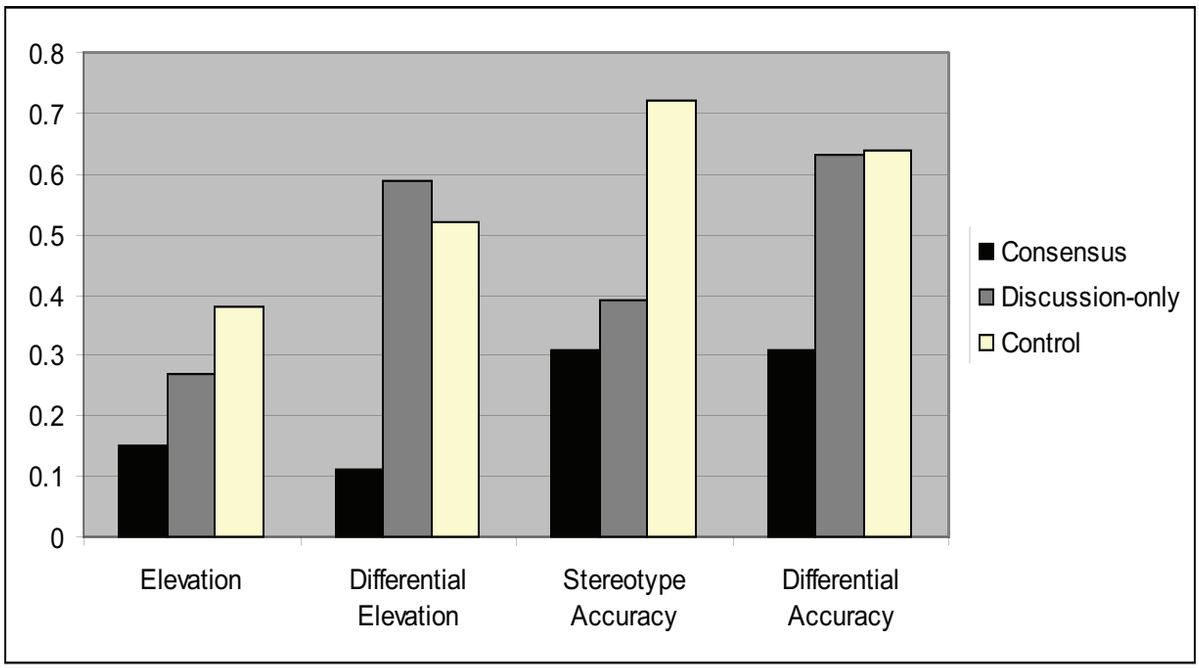
In addition to providing individual ratings for each ratee across the three performance dimensions, all participants provided an overall summary rating for each ratee and each performance dimension (nine rating scores per participant) after they viewed the video. The summary rating requirement ensured that all participants provided ratings for each of the ratees across each of the performance dimensions and did not omit any rating scores.

Following the second set of summary rating score completion, the participants were debriefed and adjourned.

Data Analysis

To test the hypothesis that participants in the consensus condition would demonstrate a significantly greater degree of accuracy in their rating scores than either the control or discussion-only conditions, we conducted a series of four one-way analysis of covariance, or ANCOVA, calculations to examine the four Cronbach (1955) measures of accuracy across experimental conditions and assess any potential covariance that may have been contributed by the initial rating score set (i.e., the pre-test) across the three experimental conditions, possibly decreasing confidence and support for any significant results produced from the final rating score set (i.e., the post-test). While the use of random assignment reduces the likelihood of a covariate impacting any causal relationship findings, inclusion of ANCOVA data

Fig. 1: Variance between Group Rating Score and True Score Means



that further reduces the likelihood of covariate influence may provide an even greater degree of support for any significant findings obtained regarding the extent to which the treatment (i.e., consensus requirement) affects the outcome (i.e., rating accuracy).

RESULTS

The following results demonstrated a significantly greater degree of rating accuracy for the consensus condition across all four accuracy indexes, thus providing support for this hypothesis (Fig. 1). The ANCOVA results for each of the four accuracy indexes are presented in Tables 1-8.

Elevation: Results indicated a significant difference between posttest scores across all three experimental conditions, F (2, 74) = 3.491, p<.05, $\acute{\eta}^2$ = 0.060, for elevation (E), which is described as the average level of rater accuracy across all ratees and performance dimensions (Table 1). Consistent with the hypothesis, the consensus rater condition demonstrated a significantly greater degree of rating accuracy than either the control condition or the discussion-only condition, as demonstrated through pairwise comparisons (Table 2).

Differential Elevation: Results indicated a significant difference between posttest scores across all three experimental conditions, F (2, 76) = 2.812, p<.05, $\acute{\eta}^2$ = 0.051, for differential elevation (DE), which is described as the degree of rater accuracy on a ratee across all performance dimensions (Table 3). Consistent with the hypothesis, the consensus rater condition demonstrated a significantly greater degree of rating accuracy than either the control condition or the discussion-only condition, as demonstrated through pairwise comparisons (Table 4).

Stereotype Accuracy: Results indicated a significant difference between posttest scores across all three experimental conditions, F(2, 74) = 2.377, p<.05, $\acute{\eta}^2$ = 0.043, for stereotype accuracy (SA), which is described as the degree of rater accuracy for each performance dimension across all ratees (Table 5). Consistent with the hypothesis, the consensus rater condition demonstrated a significantly greater degree of rating accuracy than either the control condition or the discussion-only condition, as demonstrated through pairwise comparisons (Table 6).

Table 1: Analysis of Covariance Summary Table-Elevation

Source	SS	df	MS	F	p	□²
Between Treatments	13.074	2	6.537	3.491	<.05	.060
Error	150.987	71	2.068			
Total	255.349	74				

Table 2: Results of LSD Pairwise Comparisons for Group Accuracy - Elevation

	Adjusted Mean Difference in Posttest Score (*b* – *a*)		
	b		
Condition (*a*)	**Condition 1**	**Condition 2**	**Condition 3**
Condition 1 *Control*			
Condition 2 *Discussion-only*	0.290*		
Condition 3 *Consensus*	0.445*	0.155*	

*p<.05

Table 3: Analysis of Covariance Summary Table-Differential Elevation

Source	SS	df	MS	F	p	□²
Between Treatments	8.022	2	4.011	2.812	<.05	.051
Error	399.611	73	5.474			
Total	502.093	76				

Table 4: Results of LSD Pairwise Comparisons for Group Accuracy – Differential Elevation

	Adjusted Mean Difference in Posttest Score (*b* – *a*)		
	b		
Condition (*a*)	**Condition 1**	**Condition 2**	**Condition 3**
Condition 1 *Control*			
Condition 2 *Discussion-only*	0.010		
Condition 3 *Consensus*	0.370*	0.380*	

*p<.05

Table 5: Analysis of Covariance Summary Table – Stereotype Accuracy

Source	SS	df	MS	F	p	$\square^2$
Between Treatments	6.441	2	3.221	2.377	<.05	.043
Error	481.875	71	9.601			
Total	518.882	74				

Table 6: Results of LSD Pairwise Comparisons for Group Accuracy – Stereotype Accuracy

	Adjusted Mean Difference in Posttest Score (*b* – *a*)		
	b		
Condition (*a*)	**Condition 1**	**Condition 2**	**Condition 3**
Condition 1 *Control*			
Condition 2 *Discussion-only*	0.335*		
Condition 3 *Consensus*	0.380*	0.045	

*p<.05

Differential Accuracy: Results indicated a significant difference between posttest scores across all three experimental conditions, $F(2, 75) = 4.801$, $p<.05$, $\acute{\eta}^2 = 0.791$, for differential accuracy (DA), which is described as the degree of rater accuracy on a specific ratee and a specific performance dimension (Table 7). Consistent with the hypothesis, the consensus rater condition demonstrated a significantly greater degree of rating accuracy than either the control condition or the discussion-only condition, as demonstrated through pairwise comparisons (Table 8).

Overall, these findings provide support for the hypothesis that the raters in the consensus condition would demonstrate a greater degree of rating accuracy than either the raters in the control condition or the raters in the discussion-only condition. The participants in the consensus condition provided significantly more accurate ratings, thus resulting in greater accuracy across all of the four Cronbach (1955) measures than the participants in the control condition or discussion-only condition.

Impact of Initial Rating Scores as a Covariate: Results indicated that the initial rating scores, treated as a pretest, did not impact the results of the posttest scores, and therefore did not act as a covariate in any of the three rater conditions. A parameter estimates analysis produced no significant *b* coefficients for either the control, discussion-only, or consensus conditions. Results obtained from Levene's Test produced no significant values for any of the experimental conditions, suggesting that the group variances were equal, or that they demonstrated homogeneity of variance. These findings suggest that the initial rating score set did not act as a covariate for any of the three experimental conditions. The results obtained do not demonstrate any significant

table 7: Analysis of Covariance Summary Table – Differential Accuracy

Source	SS	df	MS	F	p	$\square^2$
Between Treatments	37.007	2	18.503	4.801	<.05	.791
Error	336.002	72	6.938			
Total	538.105	75				

Table 8: Results of LSD Pairwise Comparisons for Group Accuracy – Differential Accuracy

	Adjusted Mean Difference in Posttest Score (*b* – *a*)		
	b		
Condition (*a*)	**Condition 1**	**Condition 2**	**Condition 3**
Condition 1 *Control*			
Condition 2 *Discussion-only*	0.025		
Condition 3 *Consensus*	0.205*	0.180*	

*p<.05

differences after partialling out the effect that the covariate may have had on the treatment outcome. It can be inferred that the significant findings previously explained were due to the manipulation of the experimental condition and these findings were not likely impacted by a covariate or any selection issues.

Discussion

Trends in Research and Practice

Performance management is a multifaceted domain, and a wide array of focal points, or constructs, are found within each of the three distinct areas of this domain: assessment, feedback, and reaction/outcomes. Research, regardless of focus, should strive to provide support for a common underlying theme: the improvement of an organisation's performance management strategy for greater utility and relevance to business objectives. Organisations have been increasingly required to contend with such issues as economic turmoil, rapid industry changes, and competitive pressures, and these issues will continue to present a challenge to organisations as they strive for profitability and success.

Organisations are using employee performance information in myriad ways, including strategic decisions for workforce planning and staffing, compensation, and training and employee development. Employers may compare incumbent performance against a set standard to determine the extent of any knowledge/skills deficits that may necessitate targeted training and development activities. Employers may also compare the performance of two or more incumbents with each other to determine appropriate allocation of merit increase and bonus distribution. From a more longitudinal perspective, employers may also examine incumbent performance information collected and recorded over a period of time to examine trends and changes in performance, necessary for high-stakes workforce planning and restructuring decisions such as employee transfer, termination, promotion, and succession planning.

Sound organisational decisions require the highest possible degree of accuracy in the performance information collected and analyzed. Moreover, these decisions should be directly aligned with short and long-term organisational objectives in order to make the most significant and positive impact, thus further emphasizing the importance of accuracy in employee performance information that affects every level of organisational strategy, from the day-to-day task delegation to long-range business goals.

Explanation of the Results

The findings obtained in the present research demonstrated the superior accuracy in the rating scores provided by the participants in the consensus condition compared to the rating scores provided by the participants in either the discussion-only or control conditions. Moreover, the consensus condition demonstrated significantly higher accuracy in all four of the Cronbach (1955) measures: elevation (E), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA). It can be inferred from this finding that rater consensus has the potential for improving all four types of accuracy, each contributing a unique and

crucial facet of incumbent performance across various performance dimensions.

These four unique types of rating accuracy have different effects on different organisational decisions, and each clearly has a place in an organisation's workforce strategy (Murphy, Garcia, Kerkar, Martin & Balzer, 1982). Elevation is an accuracy index appropriate for measuring overall performance across incumbents and performance dimensions, which may be an important metric for the assessment of team and/or departmental performance. Differential elevation is an accuracy index appropriate for distinguishing differing levels of performance between ratees, which may be appropriate for compensation decisions such as merit increase or bonus distribution. Stereotype accuracy is an accuracy index appropriate for measuring individual ratee performance across dimensions, necessary for personnel decisions such as probation, termination, or promotion. Differential accuracy is an accuracy index appropriate for the most fine-grained level of measurement of individual ratee performance on a specific performance dimension, appropriate for identification of areas in which targeted skills training, coaching, and other developmental activities may be warranted.

From these descriptions and potential applications, it is apparent that the value of each accuracy index is dependent on the organisational need and situation. Further research examining the construct of rating accuracy should take into consideration the type of rating accuracy of interest, and incorporate one or more of the aforementioned Cronbach (1955) measures accordingly.

Research Limitations

The present study may possess some limitations consistent with related research (Murphy *et al*., 1982, Roch, 2006). While important significant findings were discovered, it would be judicious to explore alternative research designs and methods for attending to any limitations evident in existing research studies.

Generalisation Issues

A limitation of the present study is in its strength of generalisation. While every effort was made to ensure a more robust sample of between 100-125 participants, the final sample was slightly under 100 (n=96) after some degree of attrition between the recruiting, scheduling and actual conducting of the sessions. Though data findings yielded partial support for the hypothesis, statistical strength and confidence in causal inference may be limited due to the smaller sample size. It will be necessary to continue this line of inquiry with a follow-up study comprised of the originally intended sample size to yield stronger statistical power and effect size. Additionally, researchers that attempt to make generalisations to applied settings from findings produced by studies in laboratory settings should do so cautiously, particularly with such organisationally-relevant constructs as employee performance ratings. Much of the research in this domain (Smither *et al*., 1989; Roach & Gupta, 1992; Mero & Motowidlo, 1995; Roch, 2006) has been conducted in controlled laboratory settings using students and other non-professionals as participants. While this is an acceptable practice in the research community, applied organisational research must also focus on replication and generalisation to field settings. It will be necessary to conduct follow-up research incorporating design elements more closely aligned with contextual factors in an applied setting, in an effort to replicate significant findings obtained in experimental settings with a greater degree of external validity.

Future Research Directions

The present study provides support for not only the use of multiple raters for providing performance ratings, but the integration of rater judgment in a consensus decision regarding incumbent performance across several dimensions. This rater consensus may be beneficial in the achievement of a higher degree of performance rating accuracy than perhaps a single rater or multiple raters providing independent ratings could provide.

While the present study involved the collection of two sets of rating score data from the participant groups, this was done solely for methodological consistency across experimental conditions. Because the research design required that the participants in the consensus condition independently complete individual ratings prior to convening as a group to reach consensus on a collective final set of ratings, it was necessary for the participants in the control and discussion-only conditions to also complete two sets of ratings for greater experimental control. However, the initial rating score data were not examined for the purpose of the present study and were solely treated as a covariate. Future research should explore the changes observed in rating scores from the initial rating set to the final rating set for all three experimental conditions, including the direction of the changes (i.e., giving a ratee better or worse performance ratings), and the extent to which the final rating score set moved closer to the true scores or further away. Consensus decision making involves input from different sources, and mutual agreement on a final outcome. Group members must take the time to discuss their input, including areas of commonality and difference, until they arrive at an outcome in which all members can agree upon and support (Roch, 2006). Therefore, the member-specific (e.g., personality differences) and contextual factors (e.g., complexity of

decision, time constraints) that occur during this process should be a focal point for future research, perhaps studied as potential moderators.

Conclusions and Implications for Research and Practice

The present study has several important implications for not only the research community in terms of support for existing findings and examination of different variables, but also for the practitioner community in terms of attempting to answer important questions and shed light on facets of the performance management process that may provide relevant insight for organisations interested in developing a new performance management strategy or improving upon an existing strategy. Performance management research should maintain a focus on and alignment with current organisational needs and projected trends. The present study was designed so as to include certain elements that are typically present in an applied organisational setting but are rarely studied empirically, and focused on several key areas of importance in the organisational context: performance evaluation using multiple raters who are required to reach consensus on rating decisions, evaluation of multiple performance dimensions simultaneously, and evaluation of multiple ratees who may be performing job tasks collectively but must be evaluated individually.

Multi-rater Consensus

Support for the finding that a consensus-driven performance rating decision provides superior accuracy is particularly important when compared with accuracy of performance ratings derived from discussion without consensus. Research (Roch, 2006) that has examined the differences in rating accuracy between a consensus approach and a discussion-only approach has yielded similar findings, specifically that a consensus-based performance rating is more accurate than a discussion-only performance rating, which is no more accurate than a control condition in which neither discussion nor consensus occurs. It can be inferred from this finding that certain processes exist within the act of reaching consensus that may enable greater accuracy of the final decision. These processes should be explored further at both the individual ratee level and the group level. This finding may be a novel concept for an organisation relying on traditional methods for performance evaluation, such as a single rater approach (e.g., a supervisor), or a multi-rater approach in which the ratings are completed by individual raters (e.g., supervisor, customer, co-worker, subordinate) and submitted independently.

Evaluation of Multiple Ratees

In most organisations, it is usually the supervisors who are responsible for evaluating incumbent performance and conducting performance appraisals. Supervisors often have several direct reports (e.g., incumbents within a specific department or shift) who work in concert with one another. While each incumbent may be working on an independent job task, there is commonly some amount of interaction with others.

Additionally, incumbents may spend a portion of their time working on teams in which there are shared tasks requiring a high degree of interdependence, interaction, cooperation, and accountability. As various teams continue to be positioned within the organisational structure to meet specific objectives, there will be an ongoing need to evaluate team performance. Assessment of team performance is valuable to a variety of organisational stakeholders, including team leaders, senior management, and external stakeholders, in terms of evaluating team member's contributions to the collective effort as well as evaluating the extent to which the team is achieving established goals effectively and expeditiously (London, 2007). Therefore, supervisors and others responsible for performance evaluation must be able to evaluate the performance of individual incumbents in the context of such interaction. To assume that a supervisor or rater will have ample opportunities to observe and evaluate an incumbent's performance while working alone would be unrealistic. The present study incorporated the multiple ratee element in an effort to be inclusive of this real-world contextual factor present in organisations. Future research should incorporate more team-based scenarios, in which several ratees are interacting and working together towards a common goal, for rating the performance of not only the individual team members and the quality of their respective contributions, but also perhaps the team as a collective entity.

Evaluation of Multiple Performance Dimensions

In a similar manner to the evaluation of multiple incumbents simultaneously, supervisors charged with evaluating the performance of their direct reports often must observe and evaluate incumbents on several individual performance dimensions that occur concurrently. For example, a retail employee performing a customer sales transaction might be demonstrating several important performance dimensions at once depending on the type and complexity of the transaction, including verbal communication, mathematical ability, detail orientation, cooperation, and problem solving. Therefore, a supervisor responsible for performance evaluations must be able to identify the appropriate performance dimensions based on observed incumbent behaviours and evaluate

each performance dimension individually despite their simultaneous occurrence. To assume that a supervisor will have opportunities to observe and evaluate incumbents on performance dimensions that occur independently is not only unrealistic, but even more unrealistic than the assumption of evaluating incumbent performance while working alone. The present study included the multiple performance dimension element in an effort to also be inclusive of this real-world contextual factor clearly demonstrated in the majority of jobs in today's organisations.

In the current organisational climate, jobs are highly complex and projects, tasks, and goals cross the traditional boundaries of department, function, and organisational level (London, 2007). The emphasis on teams, including cross-functional teams, virtual teams, production teams, and management teams, will continue to exist and expand in organisations in response to the need for improved workforce agility, responsiveness, and competency to meet goals of higher speed and complexity.

No longer is the traditional performance evaluation model in which a supervisor is solely aware and in full understanding of an incumbent's performance the norm; the performance evaluation process must become more dynamic and adaptable in response to increasingly complex jobs, greater incumbent interaction and collaboration, and constantly shifting organisational objectives. The present research study attempted to provide valuable insight into the extent to which a multiple rater consensus performance evaluation model may improve the accuracy of performance ratings in the context of two very relevant organisational factors, multiple ratees and multiple performance dimensions. The insight offered through this research study is not only relevant to all organisations today, but may prove to be a feasible approach in improving upon the accuracy of the performance management strategy, an extremely beneficial initiative at not only the incumbent level but for the organisation as a whole.

References

Balzer, W. K., & Sulsky, L. M. (1990). Performance appraisal effectiveness. *Psychology in Organizations: Integrating Science and Practice*, 133-156.

Barnes-Farrell, J. L., & Lynch, A. M. (2003). Performance appraisal and feedback programs. In J.E. Edwards, J.C. Scott, & N.S. Raju (Eds.), *The Human Resources Program-Evaluation Handbook*. Thousand Oaks, CA: Sage Publications, 155-176.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, 20, 238-252.

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

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychological Bulletin*, 52, 177-193.

DeNisi, A. S., & Peters, L. H. (1996). Organization of information in memory and the performance appraisal process: Evidence from the field. *Journal of Applied Psychology*, 81(6), 717-737.

Guion, R. M. (1965). *Personnel Testing*. New York: McGraw-Hill.

Ilgen, D. R., & Favero, J. L. (1985). Limits in generalization from psychological research to performance appraisal processes. *Academy of Management Review*, 10, 311-321.

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

Landy, F. J., & Farr, J. L. (1983). *The Measurement of Work Performance*. New York: Academic Press.

London, M. (2007). Performance appraisal for groups: Models and methods for assessing group processes and outcomes for development and evaluation. *Consulting Psychology Journal: Practice and Research*, 59(3), 175-188.

London, M., Mone, E. M., & Scott, J. C. (Winter 2004). Performance management and assessment: Methods for improved rater accuracy and employee goal setting. *Human Resource Management*, 43(4), 319-336.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology*, 70, 66-71.

McGourty, J. (2001). *The Team Developer: An Assessment and Skill Building Program Instructors Resource Guide*. Hoboken: John Wiley & Sons, Inc.

Martell, R. F., & Borg, M. R. (1993). A comparison of the behavioral rating accuracy of groups and individuals. *Journal of Applied Psychology*, 78(1), 43-50.

Martell, R. F., & Guzzo, R. A. (1991). The dynamics of implicit theories of work group performance: When and how do they operate? *Organizational Behavior and Human Decision Processes*, 50, 51-74.

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology*, 80(4), 517-524.

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.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, 74, 619-624.

Murphy, K. R., Balzer, W. K., Kellam, K. L., & Armstrong, J. G. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. *Journal of Educational Psychology*, 76, 45-54.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology*, 67, 320-325.

Nieman-Gonder, J. (2006). The effect of feedback and development planning on subsequent behavioral ratings and objective group performance. Unpublished doctoral dissertation, Hofstra University, Hempstead, NY.

Roach, D. W., & Gupta, N. (1992). A realistic simulation for assessing the relationships among components of rating accuracy. *Journal of Applied Psychology*, 77(2), 196-200.

Roch, S. G. (2006). Discussion and consensus in rater groups: Implications for behavioral and rating accuracy. *Human Performance*, 19(2), 91-115.

Roch, S. G. (2006). Benefits of rater teams: Role of consensus and rater motivation. Poster presented at the 2006 annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Roch, S. G., & Woehr, D. J. (1997, August). The effect of rater motivation on the accuracy of performance evaluation: An NPI approach. Paper presented at the annual meeting of the American Psychological Association, Chicago, IL.

Salvemini, N. J., Reilly, R. R., & Smither, J. W. (1993). The influence of rater motivation on assimilation effects and accuracy in performance ratings. *Organizational Behavior and Human Decision Processes*, 55, 41-60.

Smither, J. W., Barry, S. R., & Reilly, R. R. (1989). An investigation of the validity of expert true score estimates in appraisal research. *Journal of Applied Psychology*, 74(1), 143-151.

Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58, 33-66.

Smither, J. W., & Reilly, R. R. (1987). True intercorrelation among job components, time delay in rating, and rater intelligence as determinants of accuracy in performance ratings. *Organizational Behavior and Human Decision Processes*, 40, 369-391.

Sulsky, L. M., & Balzer, W. K. (1986, May). The behavioral diary format: Increasing rating accuracy through consideration of rater cognitive processes. Paper presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, 73(3), 497-506.

Tetlock, P. E. (1985). Accountability: The neglected social context of judgment and choice. In B.M. Staw & L. Cummings (Eds.), *Research in Organizational Behavior* (Vol. 7, 297-332). Greenwich, CT: JAI Press.

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, 79(4), 525-534.