Int. J. Manag. Bus. Res., 3 (2), 121-134, Spring 2013

Self – Others Rating Discrepancy of Task and Contextual Performance

[*] ***D. Wahyu Ariani***

Department of Management, Atma Jaya University, Yogyakarta, Indonesia

ABSTRACT:
This research compared ratings of task performance and contextual performance from three different sources: self, peer, and supervisor. Participants were service industry employees in the service industries in Yogyakarta, Indonesia. A Sample of 146 employees and 40 supervisors from the service industries provided ratings of task performance and contextual performance. The results indicated that there were significant differences in the mean ratings across the two sources. Self-ratings and peer-ratings of task and contextual performance are not significantly different, but self-rating and supervisor-ratings of task and contextual performance are significantly different. Peer-ratings are significantly different from supervisor ratings of task performance, but not significantly different of contextual performance. Using MTMM matrix, there is a convergence for self-rating and peer-ratings of task and contextual performance. I also find strong method effects, indicating that ratings from different sources provide different information. Using raters from different levels may also help to develop consensus, eliminate bias, and perhaps in turn lead to general acceptance by ratee. Practitioner considering the use of self-rating should be aware that there is liable to be much disagreement.

Keywords: *Task performance, Contextual performance, Self-rating, Peer-rating, Supervisor-rating*

INTRODUCTION

Multisource performance rating systems have become increasingly popular in the recent years. Common source includes supervisors, peers, and self-ratings. Subordinate rating would show low correlation with all other sources (Conway and Huffcutt, 1997). This is mainly due to the fact that subordinates are likely to observe a relatively small and different portion of their managers' job performance, relative to the amount of observation by supervisor or peers. The assumption underlying the use of multiple source ratings in 360^0 program should be examined and evaluated empirically. One of the assumptions is that ratings from different organizational levels provide different, relatively unique perspectives (Borman, 1997). Research supporting this assumption is that interrater agreement within organizational level has generally been found to be higher than across level. One issue that is particularly important is the congruence between self-ratings and other ratings (e.g., ratings from source such as supervisors, peers, subordinates, and customers) (Mersman and Donaldson, 2000). These ratings affect how results from 360^0 feedback are interpreted and presented to participants.

The extent of rater congruence in multi rater systems is of practical importance because it affects how results are interpreted and presented to participants. Self-ratings tend to be more controversial: they have been identified as lenient and restricted in range, possessing halo, and not having construct validity (McEnery and Blanchard, 1999). Convergence between self

***Corresponding Author, Email:** *dwariani@gmail.com*

and other ratings of performance may be an indicator for convergent validity, leniency bias, self-awareness, or accuracy. Yamarino and Atwater (1997) defined accurate ratings that are in agreement, and accurate estimators as those who rate themselves in alignment with how others rate them. It is consistent with what Bozeman (1997) said the interrater agreement than convergence leads to reliability which subsequently leads to validity. It is typically believed that lack of agreement indicates invalid ratings.

The primary problem with self-ratings is that they frequently disagree with supervisor ratings and they differ in the expected direction, the employees rate themselves higher than does the supervisor (Tsui and Barry, 1986). The reasons for discrepancies are informational differences about what is to be performed and how to be performed, different schemes associated with employee performance and psychological defense by the employee about their performance. Peer ratings are unfortunately highly unreliable. Shore, Shore, and Thornton (1992) found that peer ratings were superior in predicting performance than were self-evaluation. Peer would obviously interact with an employee in a much different manner than supervisors or subordinates.

According to classical measurement theory, validity and reliability are related. The relationship between validity and reliability is not reciprocal (Kasten and Nevo, 2008). Validity implies a minimum of reliability, but not reverses. Interrater reliability is considered as an important feature of performance appraisal quality. In the context of performance measurement, it assesses the extent to which different raters are consistent in appraising the performance of several individuals. Interrater reliability is used as the main reliability estimate for the correction of validity coefficients when the criterion is that of job performance ratings. Schmidt, Viswesvaran, and Ones (2000) asserted that interrater reliability is the only appropriate reliability index for this purpose. The most popular measure of inter rater reliability is the correlations between raters.

Numerous advantages of using multiple raters have been cited, for example, enhanced ability to observe and measure various job facets, greater reliability, fairness, ratee acceptance, and improve defensibility of the performance appraisal program from a legal standpoint (Harris and Schaubroeck, 1988). Comparing different rating sources (especially supervisors and peers) has been a frequent research topic. Both Harris and Schaubroeck (1988) and Conway and Huffcutt (1997) found relatively low correlations between self and other raters. Disagreement between self-ratings and other ratings can be in either two directions: self-ratings greater than other ratings and self-ratings less than other ratings (Mersman and Donaldson, 2000).

Some researches have observed relatively high peer-supervisor correlations, but others have found much lower correlations between peer and supervisor performance ratings (Harris and Schaubroeck, 1988). This is because different raters may observe different dimensions of performance or have different definitions of effective performance. Yu and Murphy (1993) found self-ratings show low to moderate correlations with ratings obtained from supervisor and peers. Self-ratings are actually higher or more lenient than ratings obtained from supervisor or peers (Harris and Schaubroeck, 1988; Yu and Murphy,1993; Khalid and Ali, 2005). Previous researches examined task performance has demonstrated a lack of agreement in performance ratings obtained from different ratings sources. Multisource issues in performance appraisal make more reliability rating, better performance information, and greater performance improvement. Convergence between self and other ratings of performance may be an indicator for convergent validity, leniency bias, self-awareness, and accuracy (Mersman and Donaldson, 2000).

Several researchers have suggested that job performance relates to two distinct sets of behavior, those that are defined in the formal job description and those that are defined by the organization's social context (Kline and Sulsky, 2009). We have two kinds of performance, in-role performance or task performance and extra-role performance or contextual performance or organizational citizenship behavior (OCB). The notion of contextual performance is important to fully describe the criterion domain of job performance (Borman and Motowidlo, 1997). Contextual performance is behavior that

contributes to the organizational, social, and psychological environment in accomplishing goals. Contextual behaviors include such behaviors as volunteering, helping, and endorsing organizational objectives and have been shown to be an important aspect of effective performance.

Frontline service employees are often the primary customers' contacts. Service employee performance can play a key role in affecting customer outcomes (Netemeyer and Maxham III, 2007). There is little evidence as how to best gauge service employee performance particularly in a service recovery context. Should employee rate themselves? Should supervisors act as the primary source of service employee performance ratings? Should coworkers set as raters of service employee performance ratings? Should some combination of three be used? Further, should just in-role performance be assessed and rewarded, or should performance beyond in-role required be assessed and rewarded as well?

Studies in rater agreement typically use task performance as the measure on which rating convergence is assessed. The extent of rater agreement using measure of contextual performance has not been adequately addressed in research to date. Moreover, the question of whether or not convergent differ according to performance dimension - task or contextual performance has not been fully explored. The aim of this study is to examine the comparability of ratings of task performance and contextual performance provided by different ratings sources. It will be done through examining the correlations between self-rating and other ratings (supervisor-ratings and peer-ratings) and the correlations between supervisor and peer-ratings as the performance appraisal in addition to task performance and contextual performance and by examining the differences between two raters: self-peer ratings, self-supervisor ratings, and peer-supervisor ratings as well.

Literature Review and Hypotheses

Researchers have traditionally thought of job performance in terms of what Borman and Motowidlo (1997) considered "task performance"—that is, employee effectiveness with regard to those activities that contribute to their organization's "technical core." Of late, several researchers (e.g., Rotundo and Sackett, 2002) have speculated that overall job performance is a function not only of task performance but also of "contextual" behavior such as OCB. Job performance, or "the set of behaviors that are relevant to the goals of the organization or the organizational unit in which a person works", remains a primary concern for organizational behavior researchers; indeed, it has been suggested that improving job performance is one of, if not the primary purposes for organizational researchers (Viswesvaran, Ones, & Hough, 2001). The fascination job performance as a topic holds for both researchers and managers lies largely in the importance of such behaviors to the organization: job performance has been shown to relate to an organization's profit, effectiveness, and survival (Motowidlo et al., 1997). Peer and self-ratings may be more useful in development for human resource decisions, because more candid ratings may be elicited (McEnery and Blanchard, 1999).

As opposed to task performance (performance defined by job descriptions and formally rewarded), contextual performance includes behaviors that are neither outlined for our expected of an employee (Borman, 1997). Behaviors in contextual performance are labeled OCB as individual behavior that in aggregate aids organizational effectiveness. This behavior is neither required by the individual's job description, nor directly rewarded by a formal reward system and as such can be thought of as extra-role performance or extra-role behavior. Extra-role behavior has been defined as individuals behaviors that are discretionary, not directly or explicitly recognized by the formal reward systems and in the aggregate promote the effective functioning of an organization.

Smith, Organ, and Near (1983) demonstrated that there are two factors in the extra-role performance scale, altruism and generalized compliance. Altruism describes the employee who directly and intentionally helps individuals in personal interaction. Generalized compliance refers to an impersonal form of conscientiousness manifested by adherence. Puffer (1987) found that a combined of compliance and altruism was related to need achievement as well as to satisfaction with rewards and to a perceived lack of peer competition. Studies in rater agreement

typically use task performance as the measure on which rating convergence in assessed. The extent of rater agreement using measure of contextual performance has not been adequately addressed in research to date. Performance feedback from multiple sources including self, supervisor, subordinate, peers, and customers has been shown to lead to more reliable ratings, better performance information, and greater performance improvements than traditional performance appraisal methods.

It is important to examine convergence across multiple measures of job performance include task and contextual performance domains (Mersman and Donaldson, 2000). Task performance defined by job description and formally rewarded. Contextual performance (citizenship behavior) is performance or behavior that is neither by individual's job description nor directly rewarded by a formal reward system, and as such can be thought as extra role performance. Lowering the scores of ratings by others such as supervisor and peers will occur. Supervisor ratings might be bias due to halo effect, memory distortion, and selective memory because contextual performance is so difficult to observed (Schnake, 1991).

One of the constraints in extra-role performance research is reliability and validity solely on ratings provided by immediate supervisors, peers, or by self-ratings. The use of self-rating of extra-role performance may be exposed to social desirability effect that is the tendency for individuals to inflate rating of their own performance (Schnake, 1991) and thus invite spuriously high correlations (Organ and Ryan, 1995). However, very little research has been conducted comparing ratings obtained from supervisors, peers, and self-ratings. The study by Becker and Vance (1993) found a moderate correlation between self-ratings and supervisor-ratings of extra-role performance. Allen, Barnard, Rush, and Russell (2000) found no relationship between these two sources of ratings. Harris and Schaubroeck (1988) use meta-analysis to determine the degree of correlation between self, peer, and supervisor ratings. They found that peer and supervisor ratings have higher correlation than self and supervisor ratings and self and peers ratings.

Drawing from Wherry's theory of rating (Wherry and Bartlett, 1982), summarized three primary factors that influence performance ratings: actual job performance exhibited by the ratee, perceptual and recall bias associated with the rater, and measurement error. The first and second factors are delineated as systematic error. Measurement error is characterized as unsystematic or random variance. Variance attributable to actual performance is considered as "true" variance. Variance attributable to rater bias is generally referred to as non random errors affecting the measurement of a concept. There are two major types of rater bias effects, halo errors and leniency error. Halo error refers to the tendency of raters to allow an overall impression of a ratee to influence judgment along several quasi-independent dimensions. Leniency error refers to rater's tendency to assign ratings that are generally higher (or lower) than are warranted by the ratees' actual performance. The other type of rater bias refers to effects associated with the raters' organizational perspective (self, subordinate, peers, and supervisor).

In peer evaluation, an individual's performance is evaluated by one or more of that individual's coworker, other than the individual's direct boss, subordinates or external customers. Peer appraisal is generally defined as the process by which an individual's colleges who are of more or less the same rank in the organization evaluate the performance of that individual. Conway and Lance (2010) found the argumentation about self report measurement. One side said that there is common method variance (CMV) in self report measures, including relationships between self-report variables are necessarily and routinely upwardly biased. Other reports (or other methods) are supervisor to self reports and rating sources (e.g. self and other) constitute measurement methods.

Studies in rater agreement typically use task performance as the measure on which rating convergence is assessed. The extent of rater agreement using measures of contextual performance has not been adequately addressed in research to date. The question of whether or not convergence differs according to performance dimension, task or contextual has not been fully explored (Conway, 1996). Correlations between sources were examined to help understand the value of including multiple sources of rating system. Interrater reliabilities

have implications for increasing the quality of rating systems and for correcting observed relations. Multitrait - Multimethod (MTMM) or Multitrait – Multirater (MTMR) framework was designed to facilitates inferences regarding the construct validity measures by examining the degree to which the same trait measured by different methods was related (convergent validity) and different traits were distinct from one another (discriminant validity).

According to classical measurement theory, validity and reliability are related. The relationship between validity and reliability is not reciprocal. Validity implies e minimal value of reliability, but not the reverse. The greater the error variance, the lower the validity coefficient (Kasten and Nevo, 2008). Researchers found a positive relationship between reliability and convergent validity and no linier relationship was found between reliability and discriminate validity. There are two methods that widely used to estimate the reliability of performance ratings. First, measures of internal consistency (α) can be used to estimate intrarater reliability. Second, measures agreement between raters can be used to estimate interrater reliability. Interrater reliability is the correlation between raters (Murphy and De Shon, 2000). Generalizability theory suggest that variance due to raters is probably not "true s core". Rater effects are quite distinct from random measurement errors and depending on the shorts of inferences one wishes to make about rating. Traditional view of interrater agreement said that convergence leads to reliability which subsequently leads to validity. Lack of agreement indicates invalid ratings. Accuracy rating is rating that are in agreement and accurate estimators as those who rate themselves in alignment with how others rate them.

Social style theory is defined as a particular pattern of actions that others can observe and agree on for describing a person's behavior. Three notable studies (Farh et al., 1991; Yu and Murphy, 1993; Furnham and Stringfield, 1994) examined self-ratings and corresponding supervisory ratings in a cross cultural context. Farh et al. (1991) found that in the collectivist culture, employee rated their own performance more harshly than did their bosses. Because collectivist culture emphasizes harmony in relationships, there is pressure for workers to understate individual accomplishments and exhibit personal modesty. Yu and Murphy (1993) and Furnham and Stringfield (1994) found self-ratings to be significantly higher for chinese employees when compared to their supervisor's ratings. Conway and Lance (2010) found that self reports are clearly appropriate for job satisfaction and many other private events, but for other constructs such as job characteristics or job performance, other type of measure might be appropriate or even superior. Therefore a hypothesis can be concluded as below:

H1: There is the difference between self-rating and other ratings of task and contextual performance.
H2: There are correlations between self-rating of task and contextual performance and supervisor-ratings of task and contextual performance.
H3: There are correlations between self-rating of task and contextual performance and peer-ratings of task and contextual performance.
H4: There are correlations between supervisor-ratings of task and contextual performance and peer-ratings of task and contextual performance.

RESEARCH METHOD

Sample and Procedure

The sample consisted of 146 employees (with response rate 97.33%) of 150 employees from service industries in Indonesia, especially in Yogyakarta. Of the 146 respondents, 83 were female and 63 were male. Each employee also rated one friend, so I have 146 self-evaluations and 146 peer-evaluations. On the other hand, 40 supervisors evaluation questionnaires return completely filled. Each supervisor rated three until five subordinates. Employee and supervisor throughout the service industries in Yogyakarta received pen-and-paper surveys. Respondents were assured of anonymity and completed the survey during working hours.

Measures

This research uses a questionnaire that is developed by some previous researchers by translating from and retranslating it to the original language. Each participant in the study was required to complete three measures: altruism, generalized compliance, and task performance. Questionnaires on the altruism and

generalized compliance are taken from those developed by previous researchers, such as Konovsky and Organ (1996); Williams and Anderson (1991); Farh, Podsakoff, and Organ (1990); Niehoff and Moorman (1993); Morrison, (1994). Job or task performance (in-role performance) was measured using items from Williams and Anderson (1991).

DATA ANALYSIS AND RESULTS

Reliability and Validity Analysis

To assess the reliability of the measurement items of all variables, an internal consistency check was carried out. The Cronbach alpha from the test yielded a record of 0.7827 for altruism, 0.8234 for generalized compliance, and 0.8192 for task performance, which is far above the cut-off line of reliability as recommended by Hair, Black, Babin, Anderson, and Tatham (2006). Content validity that is used to assess for the measurement instruments was done in the pre-tested stage by soliciting the expert opinions of two professors from a university who are research specialists in quantitative methodology and organizational behavior disciplines. The scale was then pre-tested on 30 respondents who were employee of service industry that have similar characteristics to the target population as suggested by Sekaran and Bougie (2010). Factor analysis (FA) was also performed on the construct under study. Factor extraction was executed and any Eigenvalue that is greater than one (1) will be adopted. To further simplify the interpretation and seek a simpler structure, the Orthogonal technique and the varimax rotation was then performed. The varimax rotated principal components factor revealed one structure factor. The factor loading recorded loading of between 0.518 and 0.850. Given all the items extracted were recorded above 0.5. With varimax rotation and factor loading of minimum 0.5 as suggested by Hair, Black, Babin, Anderson, and Tatham (2006) the results of construct validity testing are practically significant.

Descriptive Statistics and Mean Difference between Two raters of Task and Contextual Performance

Factor analysis is carried out to test construct validity. Then, with varimax rotation and factor loading the minimum of 0.5 as suggested by Hair, Black, Babin, Anderson, and Tatham (2006) is achieved as a result of construct validity test which is practically significant. Then, the items that have the construct validity with the use of factor analysis are tested for their reliability. Based on theoretical and empirical estimations all variables were hypothesized to be positively related. Means, standard deviation, and mean difference between two raters are provided in tables 1 and 2.

Table 1 shows the average and relation between variables used in the research. The table shows that the average altruism of self raters is (3.5499) lower than that of supervisors' (3.6918) and peers' (3.7290). The deviation standard of altruism of self raters is (0.54452) higher than that of supervisors' (0.53312) but lower than peers' (0.62003) which result in index rate of 6.5193 for self-evaluation, 6.9248 for supervisors' evaluation, and 6.0142 for peer's evaluation. It doesn't show the existence of leniency bias in the altruism when using the self rating. Employees tend to make the objective evaluation. The deviation standard of supervisors' evaluation is lower than self-evaluation. This is the evident that supervisor doesn't know the subordinate's altruism. The deviation standard of peers' evaluation is higher than self-evaluation. This is the evident that peers know the coworkers' altruism. Based on the t-test, self-rating and peer-ratings and self-rating and supervisor-ratings are significantly different, but peer-ratings and supervisor-ratings is not significantly different.

The average generalized compliance of self raters is (4.3740) lower than that of supervisors' (4.4795) but higher than peers' (4.2521). The deviation standard of generalized compliance of self raters is (0.51796) higher than that of supervisors' (0.50840) but lower than peers' (0.54383) which result in index rate of 8.4447 for self-evaluation, 8.8110 for supervisors' evaluation, and 7.8188 for peer's evaluation. It doesn't show the existence of leniency bias in the generalized compliance when using the self rating. The deviation standard of supervisors' evaluation is lower than self-evaluation. This is the evident that supervisor doesn't know the subordinate's generalized compliance. The deviation standard of peers' evaluation is higher

than self-evaluation. This is the evident that peers know the coworkers' generalized compliance. Based on the t-test, self-rating and peer-ratings and peer-ratings and supervisor-ratings are significantly different, but self-rating and supervisor-ratings are not significantly different.

Table 1: Means, Standard Deviations, and Mean Differences Between Two Raters of Two Dimensions of OCB (Altruism and Generalized Compliance)

Types Rater	Mean	Std. Deviation	Std. Error Of Mean	t	Sign.
Altruism – self-rating	3.5499	0.54452	0.04507		
Altruism – peer-rating	3.7290	0.62003	0.05131	3.164	0.002
Altruism – self-rating	3.5499	0.54452	0.04507		
Altruism – supervisor rating	3.6918	0.53312	0.04412	2.342	0.021
Altruism – peer-rating	3.7290	0.62003	0.05131		
Altruism – supervisor rating	3.6918	0.53312	0.04412	0.533	0.595
Generalized Compliance – self-rating	4.3740	0.51796	0.04287		
Generalized Compliance – peer-rating	4.2521	0.54383	0.04501	2.415	0.017
Generalized Compliance – self-rating	4.3740	0.51796	0.04287		
Generalized Compliance – supervisor-rating	4.4795	0.50840	0.04208	1.715	0.089
Generalized Compliance – peer-rating	4.2521	0.54383	0.04501		
Generalized Compliance – supervisor-rating	4.4795	0.50840	0.04208	3.680	0.000

Table 2: Means, Standard Deviations, and Mean Differences Between Two Raters of Task Performance and Contextual Performance

Types Rater	Mean	Std. Deviation	Std. Error Of Mean	t	Sign.
Task Performance – self-rating	3.9853	0.36824	0.03048		
Task Performance – peer-rating	4.0254	0.41120	0.03403	0.999	0.319
Task Performance – self-rating	3.9853	0.36824	0.03048		
Task Performance – supervisor rating	4.5205	0.47919	0.03966	11.246	0.000
Task Performance – peer-rating	4.0254	0.41120	0.03403		
Task Performance – supervisor rating	4.5205	0.47919	0.03966	8.959	0.000
Contextual Performance – self-rating	3.9619	0.40401	0.03344		
Contextual Performance – peer-rating	3.9903	0.48837	0.04042	0.659	0.511
Contextual Performance – self-rating	3.9619	0.40401	0.03344		
Contextual Performance – supervisor rating	4.0857	0.45357	0.03754	2.488	0.014
Contextual Performance – peer-rating	3.9903	0.48837	0.04042		
Contextual Performance – supervisor rating	4.0857	0.45357	0.03754	1.681	0.095

The table 2 shows the average task performance of self raters is (3.9853) lower than that of supervisors' (4.5205) and peers' (4.0254). The deviation standard of task performance of self raters is (0.36824) lower than that of supervisors' (0.47919) and peers' (0.41120) which result in index rate of 10.8226 for self-evaluation, 9.43363 for supervisors' evaluation, and 9.7894 for peer's evaluation. It shows the existence of leniency bias in the task performance when using the self rating. The deviation standard of supervisors' evaluation is higher than self-evaluation. This is the evident that supervisor knows the subordinate's task performance. The deviation standard of peers' evaluation is higher than self-evaluation. This is also the evident that peers know the coworkers' task performance. I can say that the job description of employees is very clear for employees, supervisors, and coworkers. Based on the t-test, self-rating and peer-ratings is not significantly different, but self-rating and supervisor-ratings and peer-ratings and supervisor-ratings are significantly different.

Contextual performance has different result of evaluation. The average contextual performance of self raters is (3.9619) lower than that of supervisors' (4.0857) and peers' (3.9903). The deviation standard of task performance of self raters is (0.40401) lower than that of supervisors' (0.45357) and peers' (0.48837) which result in index rate of 9.8064 for self-evaluation, 9.0079 for supervisors' evaluation, and 8.1706 for peer's evaluation. It shows the existence of leniency bias in the contextual performance when using the self-rating. The deviation standard of supervisors' evaluation is higher than self-evaluation. This is the evident that supervisor knows the subordinate's contextual performance. The deviation standard of peers' evaluation is higher than self-evaluation. This is also the evident that peers know the coworkers' contextual performance. I can say that the job description of employees is very clear for employees, supervisors, and coworkers. Based on the t-test, self-rating and peer-ratings and peer-ratings and supervisor-ratings are not significantly different, but self-rating and supervisor-ratings are significantly different.

Based on analyses of Table 2, hypothesis 1 is not fully supported, because self-peer ratings are not significantly different for contextual and task performance; self-supervisor ratings are not significantly different for task and contextual performance; and self-supervisor ratings are significantly different for task performance but not significantly different for contextual performance. The smaller deviation standard also shows that the individuals tend to see themselves as good. Peers' evaluation has a greater deviation standard and smaller average which shows peers are more objective in evaluating their subordinates' performance. Hypothesis 2, 3, and 4 postulated that there are relationship between self-rating and supervisor-ratings, self-rating and peer-ratings, and supervisor-ratings and peer-ratings. Table 3 shows correlations between self-rating and others-ratings of task performance and contextual performance.

Table 3: Correlations between Two Raters of Contextual Performance and Task Performance

	1	2	3	4	5	6
Contextual performance - self-rating (1)	1.000	0.021	0.328**	0.328**	0.007	0.258**
Contextual Performance – supervisor-ratings (2)		1.000	-0.056	-0.030	0.227**	-0.126
Contextual Performance – peer-ratings (3)			1.000	0.134	-0.020	0.514**
Task Performance – self-rating (4)				1.000	0.097	0.230**
Task Performance – supervisor-ratings (5)					1.000	-0.121
Task Performance – peer-ratings (6)						1.000

Notes: **correlation is significant at the 0.01 level (2-tailed)

Table 3 shows significant correlation between contextual performance and task performance in the same rater. Correlation between self-rating of contextual performance and task performance is 0.328, between peer-ratings of contextual performance and task performance is 0.514, and supervisor-ratings of contextual performance and task performance is 0.227. An examination of correlation matrix indicates that there are significant but moderate relationship between self-rating and peer-ratings of contextual performance (0.328) and significant but moderate relationship between self-ratings of task performance (0.230). That correlation matrix also shows that there are no relationship between peer and supervisor ratings of task and contextual performance and between self and supervisor ratings of task and contextual performance. Based on analyses of Table 3, hypothesis 2 is not supported, because correlations between self ratings and supervisor-ratings of task and contextual performance are not significant. Hypothesis 3 is supported, because correlations between self ratings and peer-ratings of task and contextual performance are significant. Hypothesis 4 is not supported, because correlations between peer-ratings and supervisor-ratings of task and contextual performance are not significant.

DISCUSSION

The results of the research show that measuring the contextual performance and task performance with the use of self and supervisor ratings indicates a significant gap. Peer-self correlations should be significant but low and peer-supervisor and self-supervisor correlations are not significant. The low correlation between raters is because three reason (Harris and Schaubroeck, 1988). First, an egocentric bias. Self rater and individuals with high self esteem may inflate their rating. Based on attribution theory, good performance because their own behavior and poor performance because environment factors. Second, difference in organizational level. Raters at different levels define and measure performance differently. Raters on the same level (self and peers) would provide similar ratings. Third, observational opportunity. Peers have more opportunities to observe ratees than supervisors. Previous research found that self-supervisor and peer-supervisor ratings exhibit lower correlations than self-peer ratings.

Research about peer and supervisor ratings of performance is important to theory and practice, as peers are frequently viewed as especially valuable sources of performance information (Greguras et al., 2003). In some environment, peers may be unwilling to evaluate each other critically. This is because they may feel that appraisal is their managers' job and they should protect their peers by not providing negative data about them. Researchers said that peers make uniformity bias. Peer-ratings would be to appraise their peers' achievement nearly the same as self-rating. Correlations for self, peer, and supervisor were moderated by job type (Conway and Huffcutt, 1997). Managerial jobs showed lower correlation than did nonmanagerial jobs and higher complexity of job was associated with lower correlations between sources.

The result of this study show that there is low correlation between self and peer ratings on task performance and contextual performance and there is no relationship between self and supervisor ratings and peer and supervisor ratings on task performance and contextual performance. A number studied found, an average, a low correlation between self-ratings and others ratings including supervisor and peer ratings. This is because individuals have a significantly different view of their own task performance than that held by other people. Peers have more opportunities to observe ratees and at more revealing times than do supervisors. Self-supervisor and peer-supervisor ratings exhibit for no significant correlations, but self-peer ratings has significant correlations. This explanation implies that supervisors disagree with a ratee because they have few opportunities to observe the individuals performance and the behaviors they do observe do not reflect true performance levels.

Correlations between self-ratings and peer-ratings are not only imperfect, they are not constants (Paunonen and O'Neill, 2010). They can be higher on some behavior attributes and lower on other attributes, and they can be higher in some groups of respondents a lower in other groups. If self-peer correlations are higher on more observable behaviors and lower on less observable behaviors and if the peer is to be

considered the ultimate criterion of target personality, this must mean that the target, and not the peer, rates him or herself more accurately on the more observable behaviors and less accurately on the less observable behaviors. The peer-ratings, not the self-ratings are affected by the observability of the behaviors underconsideration. Peers have a greater opportunity to observe performance than supervisors and that this increased opportunity may lead to improve evaluation (Stubblebine, 2001). Stubblebine (2001) found that for peers evaluation is usually not part of their formal job responsibilities, nor are they compensated for it.

Interrater reliability is the correlation between raters (Murphy and De Shon, 2000), but the systematic rater effects not associated with random measurement error affect job performance ratings (Kasten and Nevo, 2008). Raters are influenced by a wide range of factors others than the wish to be accurate, such as their personal relationships with the rater or their desire to motivate subordinates or peers. The higher level of self-peer agreement is due to personal relevance, while the lower level of self-peer agreement is the result of social relevance (Koestner et al., 1994). Personal relevance refers to whether a trait is central to a person's self identity (self motives) whereas social relevance refers to the perceived social value of given trait (social motives).

Yu and Murphy (1993) cited well documented differences between Eastern and Western cultures in terms of the relative emphasizes placed on individualism versus collectivism, and suggested that self-ratings were lower than ratings obtained from others. The results of this study also show that there is leniency bias of self-rating. This study suggested that leniency of self-ratings observed in western research may not be universal pattern. This study is in Eastern (in Indonesia, especially in Yogyakarta) and this study show self-rating is more lenient than other ratings. The quality of peer-ratings is very sensitive to the context in which the ratings are obtained when peer-ratings were conducted for evaluative purpose. Peer raters tended to rate each other more leniently and to assign similar ratings across ratees as well as across dimensions. In the other side, peer assessment as a source of performance appraisal has high reliability and validity (Farh et al., 1991).

In certain situation, peers may be better source of information regarding employee performance than supervisors. For instance, peers may have closer and more frequent contact with ratees than supervisors. They may be able to assess a wider range of performance dimensions or to make more precise performance distinction across ratees. The information possessed by peers concerning employee performance may in fact be more accurate than that possessed by any other rater. In some context, peer appraisals may be necessary because supervisors or administrators are not capable of accurately evaluating performance do to lack of knowledge about the individual's particular specialty area. Peer assessment as a source of performance appraisal has high reliability and validity Greguras et al., 2003). Although predictive validity and reliability of peer-ratings have been well established, the acceptability of peer ratings on the part of ratees appears to be problematic. One of the major obstacles constraining the use of peer performance appraisals appears to be the problem of ratee acceptance.

Barclay and Harland (1995) found that several contextual factors do appear to influence the acceptance of peer appraisals (although not all factors are consistent across samples): the nature of the performance dimensions being evaluated (they should be dimensions that peers have a good opportunity to observe), the number of raters (more is better than fewer), the experience of the ratee (more experienced ratees are less accepting of peer ratings), the validity of the ratings (higher is better), the bias of the rater (lower is better), and the leniency of the ratings (higher is better).

The result of this study also shows differences of self-supervisor ratings and peer-supervisor ratings of task performance, and differences of self-supervisor ratings of contextual performance. The non-convergent evaluation of the three raters is supported by some theories. Wheery's theory of rating shows the existence of three factors that influence the performance evaluation, that is the evaluation on the ratee's actual job performance, some bias on raters' perception, memory on the ratee's

performance, and miscalculation (Wheery & Bartlett, 1982). Based on the theory, the gap between self-rating and supervisor-rating is caused by perception bias towards the task performance and contextual performance. Borman's (1997) research also shows the same results. He states that there are reasons for the gaps, amongst them are: (1) raters' different perspectives; (2) raters from different perspectives see the same work aspects but give different weight; and (3) raters from different perspectives observe samples of different behaviors. Schnake (1991) said that the reasons of differences between raters are: (1) type of occupation and organization; (2) requirements of interpersonal interaction; (3) the degree of task interdependence; (4) organizational culture and work nature; (5) style of management; and (6) personal characteristics.

Evaluating performance by using self rating has some weaknesses. Among others is *true halo*, a mistake or bias in evaluating each of the work dimensions (Scullen et al., 2000). Besides, the rater's bias and mistakes result from the influence of the interaction between raters and ratees, and also the existence of leniency bias, that is to tend to overvalue or undervalue. The correlation between self and supervisor rating is considered low (see Harris and Schaubroeck, 1988; Furnham and Stringfield, 1994; Conway and Huffcutt, 1997; Nowack, 1997; Allen et al., 2000; Korsgaard et al., 2003; Suliman, 2003; Van der Heidjen and Nijhof, 2004; Khalid and Ali, 2005). According to Harris and Schaubroeck (1988), the low correlation between the two raters is caused by egocentric bias, gaps between organizational levels, and opportunity to observe. Egocentric bias is a result of the high self assurance (Baird, 1977; Conway and Huffcutt, 1997).

Rater from different source provides unique performance, relevant information to the rate and would not be captured by traditional supervisory ratings alone. Rater effects refer to two distinct source of variance in performance ratings: variance attributable to individual rater and variance attributable to rater source. Existing research has adopted two primary approaches to assess the presence and pervasiveness to source effects. Research has compared the correspondence of ratings from same source rates (e.g. self and peer) to that of different source raters (e.g. peer and supervisor). The next is that self perception and self improvement approaches also state that an individual with positive self image will regard him/herself as a good performer. The balance and inappropriateness theories state that there is a factor that influences self-evaluation: workers' self image. In the balance theory, there is a need to keep stable and consistent orientation towards oneself, others and environment. All the three theories support the research findings.

For contextual performance, self rating were different from others rating. The use of self-rating of contextual performance may be exposed to social desirability bias that is the tendency for individuals to inflate ratings of their own performance (Schnake, 1991) and invite spuriously high correlation (Organ and Ryan, 1995). This is because contextual performance consist of a great variety of behaviors, thus ratings provided by self and supervisors may not be strongly correlated. Although not specifically focused on peer-ratings, several studies investigating multisource ratings have indicated that a rather large portion of the total variance in performance ratings is attributable to systematic sources other than central ratee performance (Dierdorff and Surface, 2007). The effects of context on performance ratings are frequently depicted as source of bias within ratings variance and as nonperformance effects are often attributable to individual raters.

The primary empirical focus has been to address whether ratings from different sources (e.g., self, supervisor, and peers) measure similar performance constructs, provide unique perspectives, or are equally reliable. Murphy and Cleveland argued that interrater agreement may be a "non issue" because difference raters may be rating different aspects of performance and/ or using different information in their evaluations (Mersman and Donaldson, 2000). Different rating source offer different perspectives and this is where the utility of multisource rating lies. A very high level of convergence could indicate that additional ratings offer redundant information and that collecting them is a waste of organizational resources. Convergence is clearly not an indicator of "true score" or accuracy in all circumstances. Although inter rater agreement among raters (e.g. between self, peers, or

supervisors) may be neither desired nor expected, disagreement between self and other raises some interesting questions about the reasons behind the discrepancy.

Ratings from different sources can be considered 'method' factors with different correlational patterns between individual differences measures and various criteria as a function of rating source (Thomason et al., 2011). They conclude in a study of multisource performance ratings from each source have different psychological meaning. The use of at least two sources for criterion measures should enhance our understanding of managerial performance and our ability to predict it. Different rating source may also be involved in judgment of potentiality either formally or informally and there is evidence of the growing use of multiple performance appraisal sources, especially peers (Viswesvaran et al., 2002; 2005).

The trend toward using multisource rating systems for administrative purpose appears to be based on the assumption that such systems provide more complete and better quality information than that gathered from single rater source (Greguras et al., 2003). It is not possible to ensure that methods effects do not influence results, but it is reasonable for reviewers to expect authors to take certain steps to reduce the likelihood of common methods bias. Raters in organization cannot be treated as interchangeable forms of a rating instrument. This is because different raters observe different behaviors and have different responsibilities. The treatment of raters as interchangeable measurement instruments implies that measurement is a primary or important aspect of performance rating in organization.

Disagreement between self and others raises some interesting questions about reasons behind the discrepancy. Self-rating higher than others is because of ratees' overestimation, whereas self-rating lower than others is because of ratees' underestimations. Raters may disagree not because of errors, but because of other reasons such as seeing or remembering different things about the ratees. On the other hand, agreement between ratees might reflect other things that true performance invariance. Ratings might be influenced by such factors as ratee age, gender, or attractiveness. Interrater correlations do not estimate the reliability of job performance ratings. Low correlations between raters reflect not only error but in some cases, raters watch the ratees in wide range of work situations.

In terms of practical implications, the results suggested that self-ratings will generally show only low correlations with ratings by others. Using raters from different levels may also help to develop consensus, eliminate bias, and perhaps in turn lead to greater acceptance by ratees. Self-rating and peer-ratings were about the same in the accuracy of their behavior reports, on average. The self-rating was better at rating same behaviors whereas the peer was better at other. It suggests that each rater has some unique knowledge that is not available to the other. A number studied found, on average, a low correlation between self-ratings and other ratings, including supervisor and peer appraisals. Individuals have a significantly different view of their own job performance than that held by other people.

CONCLUSION

Self-rating of performance appraisal is more lenient or higher than ratings obtained from supervisor or peers. The lack of agreement across different sources suggests that multiple rating perspectives are necessity since the ratings of different source do vary. Low correlation does not imply a lack of validity or poor accuracy in the ratings of any one source. Differences between the rating sources regarding what job behaviors are expected and what job behaviors are considered as above and beyond expectations. Job type, job level, types of organization, and aspects of the performance appraisal system affect the degree of self-other ratings convergence. Individual characteristics influence the amount of self-other ratings agreement obtained in multirater systems. Limitation of this study is the small sample size of method respondents may limit the generalizability of my results. My respondents came from a variety of organizations as opposed to a sample drawn from a single organization.

REFERENCES

Allen, T. D., Barnard, S., Rush, Michael C. and Russell, J. E. A. (2000). Ratings of Organizational Citizenship Behavior: Does the Source Make A Difference? *Human Resource Management Review*, 10 (1), pp. 97-114.

Baird, L. S. (1977). Self and Supervisor Ratings of Performance: As Related to Self-esteem and Satisfaction with Supervision. *Academy of Management Journal*, 20 (2), pp. 291-300.

Barclay, J. H. and Harland, L. K. (1995). The Impact of Rater Competence, Rater Location, and Rating Correctability on fairness Perceptions. *Group and Organization Management*, 20 (1), pp. 39-60.

Becker, T. E. and Vance, R. J. (1993). Construct Validity of Three Types of Organizational Citizenship Behavior: An illustration of The Direct Product Model with Refinement. *Journal of Management*, 19 (3), pp. 663-682.

Borman, W. C. (1997). 360^0 Ratings : An Analysis of Assuimptions and A Research Agenda For Evaluating Their Validity. *Human Resource Management Review*, 7 (3), pp. 290 – 315.

Borman, W. C. and Motowidlo, S. J. (1997). Task Performance and Contextual Performance: The Meaning for Personnel Selection Research. *Human Performance*, 10 (2), pp. 99-109.

Bozeman, D. P. (1997). Inter Rater Agreement in Multi-source Performance Appraisal: A Commentary. *Journal of Organizational Behavior,* 18 (4), pp. 313–316.

Conway, J. M. (1996). Additional Construct Validity Evidence for the Task/Contextual Performance Distinction. *Human Performance*, 9 (4), pp. 309–329.

Conway, J. M. and Huffcutt, A. I. (1997). Psychometric Properties of Multisource Performance Ratings: A Meta-Analysis of Subordinate, Supervisor, Peer, and Self-Ratings. *Human Performance*, 10 (4), pp. 331-360.

Conway, J. M. and Lance, Ch. E. (2010). What Reviewers Should Expect for Authors Regarding Common Method Bias in Organizational Research. *Journal of Business Psychology*, 25 (3), pp. 325-334.

Dierdorff, E. C. and Surface, E. A. (2007). Placing Peer Ratings in Context: Systematic Influences Beyond Ratee Performance. *Personnel Psychology*, 60 (1), pp. 93-126.

Farh, J-L., Podsakoff, Ph. M. and Organ, D. W. (1990). Accounting for Organizational Citizenship Behavior: Leader Fairness and Task Scope versus Satisfaction. *Journal of Management*, 16 (4), pp. 705-721.

Farh, J-L., Cannella, J., Alberta, A. and Bedeian, A. G. (1991). The Impact of Purpose on Rating Quality and User Acceptance. *Group and Organization Studies*, 16 (4), pp. 367-386.

Furnham, A. and Stringfield, P. (1994). Congruence of Self and Subordinate Ratings of Managerial Practices as a Correlate of Supervisor Evaluation. *Journal of Occupational and Organizational Psychology*, 67 (1), pp. 57–67.

Greguras, G. J., Robie, Ch., Schleicher, D. J. and Goff, M. (2003). A Field Study of the Effects of Rating Purpose on the Quality of Multisource Ratings. *Personnel Psychology*, 56 (1), pp. 1-21.

Hair, J. E., Black, W. C., Babin, B. J., Anderson, R. E. and Tatham, R. L. (2006). *Multivariate Data Analysis*, 6 th ed. New Jersey: Prentice-Hall International Inc.

Harris, M. M. and Schaubroeck, J. (1988). A Meta-Analysis of Self-Supervisor, Self-Peer, and Peer-Supervisor Ratings. *Personnel Psychology*, 41 (1), pp. 43-62.

Hoffman, B., Lance, Ch. E., Bynum, B. and Gentry, W. A. (2010). Rater Source Effects Are Alive and Well After All. *Personnel Psychology*, 63 (1), pp. 119-151.

Kasten, R. and Nevo, B. (2008). Exploring the Relationship between Inter Rater Correlations and Validity of Peer Ratings. *Human Performance*, 21, pp. 180–197.

Khalid, Sh. A. and Ali, H. (2005). Self and Superior Ratings of Organizational Citizenship Behavior: Are There Differences in the Source of Ratings? *Problems and Perspectives in Management*, 4, pp. 147-153.

Kline, B., Theresa, J.B. and Sulsky, L. M. (2009). Measurement and Assessment Issues in Performance Appraisal. *Canadian Psychology*, 50 (3), pp. 161-171.

Koestner, R., Bernieri, F. and Zukerman, M. (1994). Sel-Peer Agreement as Function of Two Kinds of Trait Relevance: Personal and Social. *Social and Behavior Personality*, 22 (1), pp. 17-30.

Konovsky, M. A. and Organ, D. W. (1996). Dispositional and Contextual Determinant of Organizational Citizenship Behavior. *Journal of Organizational Behavior*, 17 (3), pp. 253-266.

Korsgaard, M., Audrey, M., Bruce, M. and Lester, S. W. (2004). The Effect of Other Orientation on Self-Supervisor Rating Agreement. *Journal of Organizational Behavior*, 25 (7), pp. 873-891.

McEnery, J. M. and Blanchard, P. N. (1999). Validity of Multiple Ratings of Business Student Performance of a Management Simulation. *Human Resource Development Quarterly*, 10 (2), pp. 155-172.

Mersman, J. L. and Donaldson, S. I. (2000). Factors Affecting the Convergence of Self-Peer Ratings on Contextual and Task Performance. *Human Performance*, 1 (3), pp. 299-322.

Morrison, E. W. (1994). Role Definition and Organizational Citizenship Behavior: The Importance of the Employee Perspective. *Academy of Management Journal*, 37 (6), pp. 1543-1567.

Motowidlo, S. J., Borman, W. C. and Schmit, M. J.

(1997). A Theory of Individual Differences in Task and Contextual Performance. *Human Performance*, 10 (2), pp. 71-83.

Murphy, K. R. and De Shon, R. (2000). Interrater Correlations Do Not Estimate the Reliability of Job Performance Ratings. *Personnel Psychology*, 53 (4), pp. 873-900.

Netemeyer, R. G. and Maxham, J. G. (2007). Employee Versus Supervivor Ratings of Performance in The Retail Customer Service Sector: Differences in Predictive Validity for Customer Outcomes. *Journal of Retailing*, 83 (1), pp. 131-145.

Niehoff, B. P. and Moorman, R. H. (1993). Justice as a Mediator of The Relationship Between Methods of Monitoring and Organizational Citizenship Behavior. *Academy of Management Journal*, 36 (3), pp. 527-556.

Nowack, K. M. (1997). Congruence between Self-Other Ratings and Assessment Center Performance. *Journal of Social Behavior and Personality*, 12 (5), pp. 145-166.

Organ, D.W. and Ryan, K. (1995). A Meta-analytic Review of Attitudinal and Dispositional Predictors of Organizational Citizenship Behavior. *Personnel Psychology*, 48 (4), pp. 775-802.

Paunonen, S. V. and O'Neill, T. A. (2010). Self-Reports, Peer Ratings and Construct Validity. *European Journal of Personality*, 24 (3), pp. 189–206.

Puffer, Sh. M. (1987). Prosocial Behavior, Non-Compliant Behavior, and Work Performance among Commission Salesperson. *Journal of Applied Psychology*, 72 (4), pp. 615-621.

Rotundo, M. and Sacket, P. R. (2002). The Relative Importance of Task, Citizenship, and Counterproductive Performance to Global Ratings of Job Performance. A Policy-Capturing Approach. *Journal of Applied Psychology*, 87 (1), pp. 66-80.

Schmidt, F. L., Viswesvaran, Ch. and Ones, D. S. (2000). Reliability Is Not Validity and Validity Is Not Reliability. *Personnel Psychology*, 53 (4), pp. 901-912.

Schnake, M. (1991). Organizational Citizenship : A Review, Proposal, Model, and Research Agenda. *Human Relations*, 44 (7), pp. 735-759.

Scullen, S. E., Mount, M. K. and Goff, M. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology*, 85 (6), pp. 956-970.

Sekaran, U. and Bougie, R. (2010). *Research Methods for Business: A Small Building Approach*, 5th ed. UK: John Wiley and Sons Ltd.

Shore, T. H., Shore, L. M. and Thornton, G. C. (1992). Validity of Self- and Peer Evaluations of Performance Dimensions in an Assessment Center. *Journal of Applied Psychology*, 77 (1), pp. 42-54.

Smith, C. A., Organ, D. W. and Near, J. P. (1983). Organizational Citizenship Behavior: Its Nature and Antecedents. *Journal of Applied Psychology*, 68 (4), pp. 653-663.

Stubblebine, P. C. (2001). Perception and Acceptance of Evaluations by Supervisor and Peers. *Current Psychology: Development, Learning, Social*, 20 (1), pp. 85-94.

Suliman, A. M.T. (2003). Self and Supervisor Ratings to Performance : Evidence from and Individualistic Culture. *Employee Relation*, 25 (4), pp. 371-388.

Thomason, S. J., Weeks, M., Bernardin, H. J. and Kane, J. (2011). The Differential Focus of Supervisors and Peers in Evaluations of Managerial Potential. *International Journal of Selection and Assessment*, 19 (1), pp. 82-97.

Tsui, A. and Barry, B. (1986). Interpersonal Affect and Rating Errors. *Academy of Management Journal*, 29 (3), pp. 586-598.

Van der, H., Beatrice, I. J. M. and Nijhof, A. H. J. (2004). The Value of Subjectivity: Problems and Prospects for 360-degree Appraisal Systems. *International Journal of Human Resource Management*, 15 (3), pp. 493-511.

Viswesvaran, Ch., Ones, D. S. and Schmidt, F. L. (1996). Comparative Analysis of The Reliability of Job Performance Ratings. *Journal of Applied Psychology*, 81 (5), pp. 557–574.

Viswesvaran, Ch., Schmidt, F. L. and Ones, D. S. (2002). The Moderating Influence of Job Performance Dimensions on Convergence of Supervisory and Peer Ratings of Job Performance: Unconfounding Construct-level Convergence and Rating. *Journal of Applied Psychology*, 87 (2), pp. 345–354.

Viswesvaran, Ch., Schmidt, F. L. and Ones, D. S. (2005). There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences. *Journal of Applied Psychology*, 90 (1), pp. 108–131.

Viswesvaran, Ch., Ones, D. S. and Hough, M. L. (2001). Do Impression Management Scales in Personality Inventories Predict Managerial Job Performance Ratings? *International Journal of Selection and Assessment*, 9 (4), pp. 277-289.

Wherry, R. J. and Bartlett, C. J. (1982). The Control Bias in Ratings: A theory of Rating. *Personnel Psychology*, 35 (3), pp. 521-551.

Williams, L. J. and Anderson, S. E. (1991). Job Satisfaction and Organizational Commitment as Predictors of Organizational Citizenship and In-Role Behaviors. *Journal of Management*, 17 (3), pp. 601-617.

Yammarino, F. J. and Atwater, Leanne E. (1997). Do Managers See Themselves as Others See Them? Implications of Self–other Rating Agreement for Human Resources Management. *Organizational Dynamics,* 25 (4), pp. 35–44.

Yu, J. and Murphy, K. R. (1993). Modesty Bias in Self-ratings of Performance: A test of the Cultural Reliability Hypothesis. *Personnel Psychology*, 46, pp. 357-363.