Journal of Industrial Psychology, 1999, 25(2), 21–27
Tydskrif vir Bedryfsielkunde, 1999, 25(2), 21–27

VARIABILITY IN MULTI-RATER COMPETENCY ASSESSMENTS

D. THERON & G. ROODT

Department of Human Resource Management,Rand Afrikaans University

ABSTRACT

The purpose of this study was to determine if significant differences exist between the multi-rater competency evaluations of employees operating within a flat organisational structure. Sixty-eight marketing employees were each evaluated by a number of raters including themselves, their managers, customers and peers. A competency questionnaire was developed by using the input of the employees who took part in the appraisal. Using paired t-tests significant differences between the various groups of raters were found. These findings and the implications thereof are discussed.

OPSOMMING

Die doel van hierdie studie was om te bepaal of daar beduidende verskille bestaan tussen die multi-beoordelaar bevoegdheidsevaluerings van werknemers wat binne 'n plat organisasiestruktuur funksioneer. Agt-en-sestig bemarkingswerknemers is elk beoordeel deur 'n aantal beoordelaars wat die werknemers self, hul bestuurders, kliënte en kollegas ingesluit het. 'n Bevoegdheidsvraelys is ontwikkel deur gebruik te maak van die insette van die werknemers wat deel geneem het aan die evaluering. Deur die gebruik van gepaarde t-toetse is gevind dat daar beduidende verskille bestaan tussen sommige van die groepe beoordelaars. Hierdie bevindinge en die implikasies daarvan word bespreek.

In a continuously changing environment, where quick response time is an essential ingredient in establishing a competitive edge, flatter organisational structures are put into place to deal with such dynamics (Bracken, 1994). Jones and Bearley (1996) suggest that a paradigm shift away from controlling and directing to coaching and participating is required in management. This paradigm shift will make way for participative management, empowerment and employee involvement. This phenomenon occurs in order to promote a quick response time to recognise and satisfy the needs of customers. Decision-making is pushed down to the lowest levels of organisations. This scenario has certain implications for performance management. Within a hierarchical structure, only the supervisor needs to evaluate the performance of subordinates. Supervisors are normally the only parties closely interacting with and supervising their subordinates. That could imply that they have the best knowledge on their subordinates' performance and as a result are the best choice of raters.

Flat organisational structures then call for changing ways of managing performance (Goodale, 1992; Jones & Bearley, 1996). Traditional performance management within a hierarchical structure is done from the top of the hierarchical structure to the bottom. In a flat organisational structure where individuals have far more autonomy, different performance evaluation methods are needed to provide a more objective measurement (Lawler, 1967; Scholtes, 1987). At present, with the evolvement of flatter organisational structures, Goodale (1992) suggests that 360° performance assessment (multi-rater assessment) is a product of participative management. According to Jones and Bearley (1996) 360° assessment refers to the use of multi-rater assessments of an individual, where all the assessment data is fed back to the particular individual.

In rapidly changing conditions, multi-rater assessment thus fulfils the need for providing the individual with a more holistic and useful set of feedback criteria, which can greatly facilitate development. Multi-rater assessment has the benefit that all employees or customers that work closely with a particular individual can participate in evaluating the performance of that particular individual. In a flat organisational structure it is not only the manager that works closely with his subordinates. Individuals function independently with a high level of interaction with their peers and customers. The measurement of the competency level of an individual can thus be done by managers, subordinates, peers, customers and by the individuals themselves. Jacobs (1989, p.36) states that, 'our perceptions of what is real and valid in the world rests on a consensus of shared beliefs... appraisal of performance is not an exception to that rule'. The rationale behind multi-rater performance evaluation is to determine the consensus of perception regarding an individuals' performance.

The problem with multi-rater performance evaluation is that the various groups of people who evaluate the ratee namely; his/her manager, peers, subordinates and customers, can differ in their assessments due to the fact that they have different working relationships with the ratee (Funder & Dobroth, 1987). Organisations with flattened structures and self-directed teams, have no need for strong control or supervision, allowing people to work independently (Jones & Bearley, 1996, p. 2). This change in organisational structure could result in limited interaction between managers and their subordinates, which in turn could have an influence on the competency assessment of their subordinates.

Previous research mainly focused on the difference between self-assessments and manager assessments. Most research indicates that employees assess themselves higher than other assessors (Bradley, 1978; Snyder, Stephan & Rosenfield, 1976). Williams and Seiler (1973) found a high average self-supervisor correlation of 0,60. Pym and Auld (1965) found a self-supervisor correlation of 0,56 across three independent studies. Baruch (1996) indicates that studies conducted in the United Kingdom and in Israel found a congruence of 0,73 and 0,81 respectively, when comparing self-performance appraisal results with direct-manager appraisals. These appraisals were used for individual employee development. These studies all indicated relatively high correlations.

However, research findings are not conclusive. Klimoski and London (1974) reported an average self-supervisor correlation of 0,05 and Ferris, Gilmore and Rowland (1985) reported a self-

Requests for reprints should be addressed to D. Theron, Department of Human Resource management, RAU, PO Box 524, Auckland Park 2006

supervisor correlation of 0,02. These correlations are very low. It is clear from research that conflicting results exist regarding the extent of self- supervisor agreement in performance assessment.

The above-mentioned studies did not take into account whether the organisational structure is hierarchical or flat in nature. The proposition can be made that in a hierarchical structure there should be a high congruence in self-supervisor performance assessments as compared to a low congruence in self-supervisor performance appraisals in a flat organisational structure. This proposition can only be made for a flat structure where individuals have a high level of independence and limited interaction with managers.

Harris and Schaubroeck (1988) examined the variability in peer and self-ratings. They found that most studies provide inconsistent findings. They suggest two reasons that might explain the inconsistency in ratings. The first reason relates to sampling error. Most of the studies are based on small sample sizes (N = 71-80). A second reason for the inconsistency lies in the nature of the job. It seems that blue-collar and service jobs have a higher consistency, based on the fact that these jobs are relatively routine and performance is well defined, compared with managerial and professional jobs with low consistency due to the fact that these jobs are not as easy to define.

Harris and Schaubroeck (1988) tried to address the above-mentioned issues in their research and found a moderate agreement between peer assessment and self-assessment (0,36), and supervisory assessment and self-assessment (0,35). A much higher average correlation was found between supervisory assessment and peer assessment (0,62).

These authors also tested for the effects of moderators. It was found that only job type had any meaningful effects. The correlations of performance assessments obtained from multi-raters were lower for professional jobs and higher in the case of service or blue/collar jobs (Harris & Schaubroeck, 1988).

Typical moderators include the following:

Egocentric Bias

Egocentric bias is based on the premise that self-ratings are biased, while other raters share a set of common perceptions (Harris & Schaubroeck, 1988, p. 44). One of the most common forms of egocentric bias, is when a rater inflates his/her rating as a result of defensiveness in order to enhance the evaluation. This leads to self-ratings having a restricted range. The correlations between assessments other than self-assessments are then considerably higher, with no range restriction.

A second version of egocentric bias assumes that ratee defensiveness may be moderated by other variables such as self-esteem. Ratees with high self-esteem may inflate their ratings, while ratees with low-esteem may not (Harris & Schaubroeck, 1988).

A third version of egocentric bias is posited by attribution theory. Raters attribute good performance to their own behaviour and bad performance to environmental factors. Observers on the other hand, attribute good performance to environmental factors and bad performance to the ratees' dispositions. This theory suggests that self-ratings will correlate poorly with other ratings. The above mentioned predicts differences in all mean ratings, and significantly higher self-ratings than either peer or supervisor ratings (Harris & Schaubroeck, 1988).

Differences in organisational level

Raters at different levels of responsibility in the organisation might weigh performance dimensions differently, based on the degree of closeness of the work relationship with the ratee (Harris & Schaubroeck, 1988).

Rater leniency

Thornton (1980, p.264) defined rater leniency as the stable tendency of certain raters to rate individuals higher in comparison with other raters. Holzbach (1978) states that leniency errors occur when ratings from different rating sources on the same ratee group are significantly different. The study done by Kane, Bernardin, Villanova and Peyrefitte (1995) confirmed the hypothesis that rater leniency is a stable rater tendency. Holzbach (1978) explored leniency effects in an analysis of supervisor, self and peer performance ratings. Leniency error represents an area of concern in that differential perceptions of performance by different sources contribute to low validity. Holzbach (1978) and Klimosky and London (1974) found self-ratings to be more lenient than either supervisor or peer ratings.

Halo Effect

Another moderator is the halo effect that is often used to explain or clarify high correlations between ratings of independent job dimensions that are directly demonstrative. The halo error is based on the premise that raters expect an employee who is performing well on one performance dimension, to be just as proficient on other performance dimensions (Holzbach, 1978). Such expectations inflate correlations over their true levels. Fahr, Cannella and Bedeian (1991) define halo error in the context of performance appraisal as a tendency of raters to allow a general impression to affect their ratings of individual dimensions, resulting in high inter-dimension correlations or low inter-dimension variance. Campbell and O'Connell (1982) emphasise that in reality there will be some level of covariation called true halo.

Holzbach (1978) measured the level of the halo effect by the magnitude of the intercorrelation among items obtained from each rating source. He found that ratings by supervisors show consistently greater halo effects than self-ratings (Klimosky, 1974). The biggest problem area with Holzbach's research is the use of a mechanistic method of choosing raters, which is removed from practical reality.

Albrecht, Glaser and Marks (1964) and Lawler (1967) found that differential perceptions and biases can be substantially reduced by the following: (1) choosing a rating group which is homogeneous; (2) by choosing raters which are familiar with the jobs and incumbents; and (3) evaluating job specific performance items. Holzbach (1978) confirms this notion in that he observed differential leniency effects where performance items were not job specific.

Purpose of ratings

Fahr, Cannella and Bedeian (1991) did a complete study of the influence of rating purpose on rating quality. They found that supervisory ratings conducted for feedback or development purposes are less prone to leniency bias.

On the other hand they found that ratings used to determine employee reward and promotability are more prone to leniency bias (Fahr, Cannelle & Bedeian, 1991; Landy & Farr, 1980). When ratings from managers, peers or customers are collected for development purposes, raters are less concerned about rating consequences and more willing to rate incumbents based on their actual performance. Under these conditions higher inter-rater agreement with ratees and more reliable ratings can be expected. Fahr, Cannella and Bedeian (1991) found a

significant correlation between self ratings and peer ratings where the purpose of the ratings were defined to be used for developmental purposes. They found peer ratings to be valid and reliable. Springer (1953) examined peer and manager ratings in a setting where the ratings were used to determine the incumbent's promotability. She found supervisory ratings to be more conservative than those of peers.

Differences in evaluation criteria

An alternative view to explain inconsistencies between self and other ratings, holds that disagreement stems from a tendency of different types of raters to base their ratings upon different aspects of job performance (Klimoski & London, 1974; Lawler, 1967; Steel & Ovalle, 1984). Different types of raters employ different types of evaluation criteria. In order to overcome this problem, performance assessment questionnaires must clearly define the evaluative criteria. Steel and Ovalle (1984) state that an attempt should be made to define the rating criteria more precisely for the rater, thus establishing a common frame of reference in the rating instructions. This might contribute to reductions in leniency errors and improved rater agreement.

In view of the discussion above, differences in the competency assessments by different rater groups that have different levels of interaction with the employee, needs to be considered. Jacobs (1989, p.32) defines competencies as observable knowledge, skills, traits and abilities used to execute a certain task or function. The value of such an analysis would lie in determining if organisational structural changes have an influence on traditional performance appraisal, specifically when the flat structure leads to reduced interaction of individuals with their managers. The practical importance of this study thus lies in determining if organisations with such flat structures should rather make use of multi-rater evaluations to ensure objective measurement of performance. A need exists to research the differences between the competency assessments of the various rater groups, using a multi-rater assessment system.

Heneman (1980) indicates that the reasons for discrepancies between manager, self, and peer ratings are not known, and that a theoretical foundation still needs to be developed to support such research. The lack of a coherent theory on multi-rater assessment clearly indicates that further research is crucial. No comprehensive study has been documented in South Africa to compare the differences between assessments of various multi-raters in a flat organisational structure setting. Consequently, the following research question is formulated.

'Do significant differences exist between multi-rater competency assessments by different rater groups within a flat organisational structure?'

The aim of this study is to determine if significant differences exist between the competency assessments of the same employees by multi-raters belonging to the following groups: managers, peers, customers and self-assessments within the context of a flat organisational structure setting.

The dramatic flattening of organisational structures poses certain challenges in terms of performance appraisal methods. It is clear that the manager is the best possible evaluator in a hierarchical structure with strict reporting systems and controls (Jones & Bearley, 1996). However, where employees operate in teams within a flat structure, and where individual responsibility is toward team members, there is limited contact between managers and their subordinates. Flattening structures might have the effect that appraisals by team members (peers) and customers should be considered in addition to managerial assessments. It implies using a more objective means of performance measurement that is in line with organisational changes. Lawler (1967) states that the multi-rater approach to performance appraisal has not received a lot of attention, but appears to have the advantage of being a more objective measure of performance.

In the section below, different hypotheses are formulated and supporting literature is quoted:

H1 A statistically significant difference exists between manager assessment and self-assessment.

The only relatively consistent finding in multi-rater research indicates that employees assess themselves consistently higher than other assessors (Bradley, 1978; Snyder, Stephan & Rosenfield, 1976; Thornton, 1980).

Baruch (1996), Pym and Auld (1965) and Williams and Seiler (1973) found relatively high correlations between self-performance appraisals and direct-manager appraisals. The sample, however, consisted of employees from government departments such as the navy, where a strict hierarchical structure exists.

In contrast, Ferris, Yates, Gilmore and Rowland (1985) and Klimoski and London (1974) found very low self-supervisor correlations. These studies, however, did not provide any indication of the moderating effect of organisational structure.

H2 A statistically significant difference exists between manager assessment and customer assessment.

No research was found that correlated the ratings of managers with customers or compared the differences between these two rater groups. It is proposed that due to a higher level of contact between the customer and the incumbent, compared with a lower level of contact between the manager and the incumbent, that manager and customer ratings will differ significantly. The above differences in ratings could be a result of flat organisational structures.

H3 A statistically significant difference exists between manager assessment and peer assessment.

Lawler (1967) indicates that manager ratings are used twice as frequently as peer ratings. However, he described peer ratings as being far more objective than manager ratings. Springer (1953) found manager ratings to be more conservative than those of peers.

Furthermore, Barclay and Harland (1995) found that peer raters were perceived as fair if they were educated and experienced. Although a lot of research has been done on fairness and accuracy of multi-rater assessments, limited research has been done on differences in assessments of multi-raters. Also, little evidence is available on the effect of organisational structure.

H4 A statistically significant difference exists between customer assessment and peer assessment.

No research was found relating customer ratings to peer ratings. However, the proposition could be made that in a company with a flat organisational structure, customer and peer ratings would be in agreement, provided both parties know the incumbent and his/her job and interacts with him/her on a frequent basis.

METHOD

Test sample

The population consists of all marketing, sales and distribution personnel (N = 128) employed by six globally operating petrochemical business units. All these business units have flat organisation structures and are operating with self-directed teams. The sampling frame included 128 marketing employees

of which 68 employees responded, which established a response rate of 52,1((Refer to Table 1).

Fifty four percent of respondents were male, while 46% were female. With regard to the representation of different language groups in the sample, English represented 28%, Afrikaans 60%, South Sotho four percent, Venda one percent and other six percent. The ages of the respondents ranged from 21 to 60, the average age being 37 years.

TABLE 1
BIOGRAPHICAL DATA OF RESPONDENTS (n = 68)

GENDER	FREQUENCY	PERCENTAGE
Male	37	54%
Female	31	46%

HOME LANGUAGE	FREQUENCY	PERCENTAGE
Afrikaans	41	60%
English	19	28%
Other	4	6%
South Sotho	3	4%
Venda	1	1%

AGE	MINIMUM	MAXIMUM	AVERAGE	STANDARD DEVIATION
	21	60	36,657	8,413

Measuring-instrument

The Marketing Competency Questionnaire that was used for this study is a custom designed assessment questionnaire. The purpose of the questionnaire is to serve as a developmental tool that can aid in identifying competency developmental areas. The dimensions measured by the questionnaire include technical competence, team skills, problem solving skills, interpersonal skills, independence, customer orientation, personal drive, planning and organising.

Each item in the questionnaire states a behavioural output that specifies high performance behaviour related to the dimension (competency). Raters are expected to evaluate both the level of importance and level of performance of a ratee in terms of the specified output. Responses were recorded on a 6-point scale of which the first number and the last number are defined. A value of one in terms of Importance refers to no importance and a value of six refers to critical importance. A value of one in terms of Performance refers to extremely poor performance and value of six refers to excellent performance. The questionnaire was developed using the repertory grid and behavioural event (critical incident) interviews as basis (Boam & Sparrow, 1992).

Reliability coefficients are reported under the heading 'Results'.

Research procedure

The questionnaires were computerised and transferred to discs. Each respondent (ratee) received his or her own evaluation disc at a group session and were given a choice of various raters. They had to choose raters with whom they had a high level of interaction. They had to choose at least one manager, one peer and one customer to evaluate them, as well as evaluating themselves. The ratees had to complete the questionnaire themselves, after which they would forward it to the various raters. On opening the questionnaire on the disc, a message box appeared with a message stating that the evaluation is for development purposes and that the raters' honest feedback would be appreciated.

In the execution of this study, an attempt was made to limit moderator influences. This was done through the following: (1) designing a job specific questionnaire relevant to the incumbents, but also developed with the input of the incumbents, (2) allowing each incumbent, together with his or her line manager, to compile a list of raters with whom they interact frequently to ensure valid feedback, (3) using real incumbents in real marketing jobs, evaluated by actual managers, peers and customers, (4) making it clear to all incumbents and their raters that the evaluation will be used to identify development areas, and not to determine compensation and (5) defining evaluative criteria for each dimension measured, to prevent raters from generalising, in order to limit halo effects (Jacobs, 1989; Klimoski & London, 1974; Lawler, 1967; Steel & Ovalle, 1984).

RESULTS

All respondents without a complete set of assessors (including a self-assessment, manager assessment, peer assessment and customer assessment) were excluded from the sample. A factor analysis procedure as suggested by Schepers (1992, pp.140–144) was conducted on the data for both the importance and performance ratings. All properly completed questionnaires that were returned were analysed statistically. The thirty items of the Marketing Competency Questionnaire were intercorrelated and subjected to a principal factor analysis. This procedure was followed for each of the rater groups. The first factor analyses identified a number of factors which differed for the rater groups. On the Importance ratings, the number of factors extracted for each of the rater groups varied form 5 to 7 factors. On the Performance rating the factors extracted varied form 6 to 8 factors. These subscores obtained, were intercorrelated and subjected to factor analyses. From this analysis, one factor was extracted including all thirty items for all the different rater groups.

Subsequently, Cronbach alpha reliability coefficients were computed for each of the scales (manager, peer, self, customer) including all thirty original items on the importance and performance ratings (Refer to Table 2).

TABLE 2
INTERNAL CONSISTENCIES (CRONBACH ALPHA) OF THE SCALES

	Consistencies on Importance Scale (all items included)	Consistencies on Performance Scale (all items included)
Manager assessment	0,943	0,941
Peer assessment	0,956	0,963
Self assessment	0,950	0,925
Customer assessment	0,954	0,954

The scales yielded alpha coefficients ranging from 0,925 to 0,963. This implies a high degree of consistency among the raters of a particular group (manager, peer, self, customer).

The second part of the data analysis was to correlate the performance ratings of the various rater groups (Refer to Table 3).

TABLE 3
INTERCORRELATIONS OF PERFORMANCE RATINGS

		Manager	Peer	Self	Customer
Manager	Pearson Correlation	1,000	0,452**	0,154	0,399**
Peer	Pearson Correlation	0,452**	1,000	0,450**	0,461**
Self	Pearson Correlation	0,154	0,450**	1,000	0,341*
Customer	Pearson Correlation	0,399**	0,461**	0,341*	1,000

Note N = 67
** – p = 0,001
* – p = 0,005

Table 3 contains the Pearson correlations for each of the rater groups on the performance ratings. The highest correlation measured was (r (df = 64; (= 0,05) = 0,461, p < 0,001) between the performance ratings of peers and customers. Moderate correlations were also found between peer and manager (r (df = 64; (= 0,05) = 0,452, p < 0,001), peer and self (r (df = 64; (= 0,05) = 0,450, p < 0,001), customer and self-ratings (r (df = 64; (= 0,05) = 0,341, p < 0,005) and managers and customers (r (df = 64; (= 0,05) = 0,399, p < 0,001).

The correlations, however, does not provide information on possible differences between rater groups. Paired t-tests for dependent samples were conducted to establish possible differences between raters (Refer to Table 4).

The first hypothesis proposed a statistically significant difference between manager assessments and self-assessments. For all the rater groups, no statistically significant differences were found on the importance rating. The Paired t-test results indicate a statistically significant difference on the performance rating between manager and self-ratings (t (df = 63; (= 0,05) = -4,054, p < 0,001, (= 0,98). A power test was done on each of the statistically significant differences, in order to determine if those differences are large enough to be of practical significance. A power (() higher than 0,8 is regarded as of high practical significance (Howell, 1997, p.226).

Secondly it was hypothesised that a statistically significant difference exists between manager and customer assessments. Again the paired t-test results indicate a statistically significant difference between manager and customer assessments (t (df = 63; (= 0,05) = -4,294, p < 0,001, (= 0,99). It was also proposed that a statistically significant difference exists between manager and peer assessments. A statistically significant difference between manager and peer assessments (t (df = 63; (= 0,05) = -3,572, p < 0,001, (= 0,94) was found.

Finally, it was proposed that a statistically significant difference would exist between the assessments of customers and peers. No statistically significant difference was found between customer and peer assessments.

The differences between manager and self-ratings, manager and customer and manager and peer ratings, with reference to the power test, can be regarded as of practical significance (Howell, 1997).

DISCUSSION

In this study an attempt was made to identify those statistically significant differences between the competency assessments of multi-raters functioning within a flat organisational structure setting. The Marketing Competency Questionnaire used, yielded high internal consistencies within the various groups of raters, both on actual performance and the level of importance of expected performance. This questionnaire can be used by other researchers in a marketing environment as a reliable instrument to assess marketing competencies. It can also be concluded that when the purpose of assessment is stated as development, higher reliability and validity of ratings are obtained (Fahr, Cannella & Bedeian, 1991).

In this study, in order to increase the reliability of the questionnaire, an attempt was made to reduce rater errors and improve rater agreement. One method applied was to clearly define the rating dimensions in order to establish a common frame of reference in the rating instructions. This frame of reference reduced the ambiguity in the rating situation. It decreased the likelihood for raters to resort to improper responses (rater leniency or halo) as a means of compensating for lack of task clarity.

It was expected that no statistically significant differences would exist between the various rater groups for the importance rating. All the rater groups, with the exception of customers were involved in the development of the questionnaire. They agreed that the questions that were finally included in the questionnaire were of critical importance for every individual functioning in a marketing environment. As a result it was expected that the rater groups would have agreement on the importance of the items.

As a first hypothesis, a statistically significant difference was proposed between manager and self-assessments. No significant correlation was found between manager and self-assessments. Using t-tests, it was determined that no significant difference exists between the ratings of managers and self-raters on the importance rating, but that a significant difference exists on the performance rating. A power test confirmed this notion.

TABLE 4
PAIRED SAMPLES T-TEST: PERFORMANCE RATING

Rater Groups	Mean of Difference Scores	Standard Deviation of Difference Scores	Std. Error Mean	t	df	Sig.(p) 2-tailed	Power δ
Manager vs. Self	–10,52	20,75	2,59	-4,054	63	0,001	0,98
Manager vs. Customer	–10,34	19,27	2,41	-4,294	63	0,001	0,99
Manager vs. Peer	–8,13	18,20	2,27	–3,572	63	0,001	0,94
Self vs. Peer	2,39	18,29	2,29	1,045	63	0,300	0,17
Customer vs. Peer	2,22	19,11	2,39	0,929	63	0,357	0,17
Self vs. Customer	0,17	19,20	2,40	0,072	63	0,943	0,17

The mean difference between manager and self-assessors presented a negative score. This indicates that the manager rated individuals more conservatively than they rated themselves. The managers also rated individuals consistently more conservative than all the other rater groups and self-assessors rated themselves consistently higher than all the other rater groups. The biggest difference exists between manager and self-assessments. Self-assessors also consistently rated themselves higher than all the other rater groups. Self-raters tend to inflate their self-assessments.

Thornton (1980, p.269) is of the opinion that self-appraisals should be used carefully. Consistent with previous research, it is indicated that self-assessments are inflated (Bradley, 1978; Snyder, Stephan & Rosenfield, 1976; Thornton, 1980). When managers are using self-appraisals with their subordinates, they should be aware that their own appraisal could differ vastly from the subordinate's self-appraisal. It seems to remain a risk to make use of self-assessment if the purpose of the assessment is anything other than development.

Previous research presented a high level of inconsistency as to the level of correlation between self-assessors and managers (Baruch, 1996; Ferris, Gilmore & Rowland, 1985; Klimoski & London, 1974; Pym & Auld, 1965; Williams & Sailer, 1973). Mabe and West (1988) found a low correlation between manager assessments and self-assessments. The trend that self-assessors consistently rate themselves higher than other raters seems to be a stable trend (Bradley, 1978; Snyder, Stephan & Rosenfield, 1976).

This difference between manager assessments and self-assessments could be explained by managers having a higher status than their subordinates, thus expecting their subordinates to conform to their own standards (Harris & Schaubroeck, 1988). It can also be attributed to the flattening of structures, where line managers do not have tight control over their subordinates. This would be relevant to with employees that have a low degree of interaction with their managers. Steel and Ovalle (1984) stated that disagreement between any manager and subordinate dyad could also be a result of the extent of openness and accuracy in the transmission of feedback between the two parties. It could also again refer back to different perceptions of role requirements. The practical implication is that performance appraisal discussions will be extremely difficult for both parties, seeing that they have different sets of perceptions pertaining to the same job function.

In practice, marketing organisations with flat organisational structures need to decide if these results should have an impact on the design of their performance appraisal systems. Considering the results, the relative importance of manager and self-assessments needs to be established when using multi-rater evaluation systems.

Secondly, it was hypothesised that a significant difference exists between manager and customer assessments. Manager and customer assessments correlated significantly, however, a significant difference was found between the performance assessments of managers and customers. A power test on this difference indicated that this difference is of practical significance.

The difference between manager and customer assessments could be the result of different relations with the respondent. In practise, in a flat marketing organisation, the respondents deal directly with the customer, only to give occasional feedback to the manager. A high level of interaction thus exists with the customer, compared to a low level of interaction with the manager. The validity of manager assessments, as compared to customer assessments can thus be questioned. Manager ratings on performance are more conservative than those of customers. No previous research could be found to shed more light on this finding.

In the third place it was hypothesised that a statistically significant difference exists between manager and peer assessments. Even though the manager and peer assessments did correlate significantly, it differed significantly on the performance rating. The ratings thus followed a similar trend, but differed significantly in score. A power test indicated that the difference between the manager and peer assessments is of practical significance.

In a flat marketing organisation, where the employees work in self-directed teams, there is also a high level of interaction with peers.

Lawler (1967) indicates that managers' ratings are used twice as frequently as peer ratings. However, he describes peer ratings as being far more objective than manager ratings. In a flat organisation, where peers have a high level of interaction, using peer assessments in addition to manager assessments, might be more valid than only using manager assessments. Springer (1953) also found manager ratings to be more conservative than those of peers. The validity of manager assessments as the only form of assessment in a flat organisation structure can be questioned considering the fact that it is significantly more conservative than peer, self and customer ratings.

Finally, it was hypothesised that a significant difference would exist between the assessments of customers and peers. No significant difference, however, were found between peer and customer assessments. The highest correlation found was between the peer and customer assessments. This could indicate that they have similar perceptions of the ratees. No significant difference was found between the assessments of peers and customers on both the importance and performance ratings. This could indicate that there exists a degree of agreement between peers and customers. No previous research was found relating the assessments of peers and customers. However, the proposition could be made that in a company with a flat organisational structure, peer and customer ratings would correlate, provided both parties know the incumbent and his/her job and interact with him/her on a frequent basis.

In considering the peer and customer ratings, it can be concluded that both the validity of only using self or manager assessments can be questioned in a flat structure. Self-ratings are inflated, and manager ratings are more conservative than the other rater groups. It is interesting that the manager formed part of all the pairs where statistically significant differences were found. Significant differences thus exist between manager and self-raters, manager and peer raters and manager and customer raters. It is even more interesting that no significant differences were found between any of the other rater combinations. The manager assessment thus seems to be rather controversial. This finding should be seriously considered in the design of performance assessment systems. It could imply that multi-rater evaluation could be valuable in an organisation with a flat structure, where the weight of the manager assessment might decrease.

It can also be concluded that multi-rater assessment is appropriate where the purpose of the evaluation is development. More objective and valid feedback on performance is obtained.

Possible limitations of this study that should be considered, are a relatively small sample size and the fact that this study is limited to marketing companies in the petro-chemical industry. Sales within the petro-chemical industry are all bulk sales, which limits the number of marketing personnel needed to sell the product. Therefore, it is impossible to obtain a large sample size in this industry. These findings can not necessarily be

generalised for another industry. With reference to the limited amount of research found, further research is needed to study the differences between multi-raters, especially examining the differences between peer and self-raters and customer and peer raters.

Furthermore, the results reported here merely show the differences of perceptions between managers, peers, customers and self-raters. Further research is necessary to confirm these findings and to identify the sources of these differences (Heneman, 1980, p.298). The number of simplified factor scores (sfs) for the various rater groups differed. A possible explanation could be that the rater groups are using different mental models to interpret the questionnaire. A further article by the same authors will explore the reasons for these differences in interpretation.

Other possible sources of differences between rater groups include personality, age and cultural influences (Ferris, Yates, Gilmore & Rowland, 1985; Pym & Auld, 1996; Sullivan & Taylor, 1991). It could be that managers' low ratings of subordinates are a result of age differences.

Further research should consider a comparison between a flat organisational structure and a hierarchical structure in companies. It would be interesting if two studies are done on each of these structures, one where the assessment is for development purposes, and the other where the assessment is done for promotion purposes or remuneration increases. It is important that sample size be increased to ensure meaningful results.

REFERENCES

Barclay, J.H. & Harland, L.K. (1995). Peer performance appraisals. *Group and Organizational Management*, 20, (1), 39–60.

Baruch, Y. (1996). Self-performance appraisal vs. direct-manager appraisal: a case of congruence. *Journal of Managerial Psychology*, 11 (6), 50 (16).

Bracken, D.W. (1994). Straight talk about multirater feedback. *Training and Development*, September, 44–51

Bradley, G.W. (1978). Self-serving biases in the attribution process: a re-examination of the fact or fiction question. *Journal of Personality and Social Psychology*, 36, 56–71.

Campbell, D.T. & O'Connell, E.J. (1982). Methods as diluting trait relationships rather than adding irrelevant systematic variance. In D. Brinberg & L. Kidder (Eds.), *New directions for methodology of social and behavioral science: Forms of validity in research.* San Francisco: Jossey-Bass.

Fahr, J., Cannella, A.A. & Bedeian, A.G. (1991). Peer ratings - the impact of purpose on rating quality and user acceptance. *Group and Organization Studies*, 16(4), 367–386.

Ferris, G.R., Yates, V.L., Gilmore, D.C. & Rowland, K.M. (1985). The influence of subordinate age on performance ratings and causal attributions. *Personnel Psychology*, 38, 545–557.

Funder, D.C. & Dobroth, K.M. (1987). Differences between traits: properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, 52, 409–418.

Goodale, J.G. (1992). Improve employee performance reviews. *Hydrocarbon Processing*, November, 231–232, 244.

Guilford, J.P. (1954). *Psychometric Methods.* New York: McGraw Hill.

Harris, M.H. & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43–62.

Heneman, H. G. (1980). Self-assessment: a critical analysis. *Personnel Psychology*, 33, 297–300.

Holzbach, R.L. (1978). Rater bias in performance ratings: superior, self- and peer ratings. *Journal of Applied Psychology*, 63 (5), 579–588.

Howell, D.C. (1997). *Statistical Methods for Psychology (4th ed).* Belmont, California: Duxbury.

Jacobs, R. (1989). Getting the measure of management competence. *Personnel Management*, June, 32–37.

Jones, E.J. & Bearley, W.L. (1996). *360 feedback: strategies, tactics and techniques for developing leaders.* Amherts: MRD.

Klimosky, R.J. & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology*, 59, 445–451.

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

Lawler, E.E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, 51,(5), 369–381.

Mabe, P.A. & West, S.G. (1982). Validity of self-evaluation of ability: a review and meta-analysis. *Journal of Applied Psychology*, 67, 280–296.

Mount, M.K., Sytsma, M.R., Hazucha, J.F. & Holt, K.E. (1997). Rater-ratee race effects in developmental performance ratings of managers. *Personnel Psychology*, 50, (1), 51 (19).

Nowack, K.M. (1993). 360 feedback: the whole story. *Training and Development* 47 (1), 69–72.

Park, B. & Judd, C.M. (1989). Agreement on initial impressions: differences due to perceivers, trait dimension, and target behaviors. *Journal of Personality and Social Psychology*, 56, 493–505.

Pym, D.L.A. & Auld, H.D. (1965). The self-rating as a measure of employee satisfactoriness. *Journal of Occupational Psychology*, 39, 103–113.

Schepers, J.M. (1992). *Toetskonstruksie, teorie en praktyk.* Johannesburg: RAU-Drukpers.

Scholtes, P.R. (1987). *Joiner – An elaboration on Deming's teachings on performance appraisal.* Madison: Joiner.

Snyder, M.L., Stephan, W.G. & Rosenfield, D. (1976). Egotism and attribution. *Journal of Personality and Social Psychology*, 33, 435–441.

Springer, D. (1953). Ratings of candidates for promotion by co-workers and supervisors. *Journal of Applied Psychology*, 37, 347–351.

Steele, R.P. & Ovalle, N.K. (1984). Self-appraisal based on supervisory feedback. *Personnel Psychology*, 37, 667–685.

Sullivan, J. & Taylor, S. (1991). A cross-cultural test of compliance-gaining theory. *Management Communication Quarterly*, 5 (2), 220–239.

Thornton, C. T. III (1980). Psychometric properties of self-appraisals of job performance. *Personnel Psychology*, 33, 263–271.

Williams, W.E. & Seiler, D.A. (1973). Relationship between measures of effort and performance. *Journal of Applied Psychology*, 57, 49–54.