SA Journal of Industrial Psychology, 2003, 29 (2), 39-48
SA Tydskrif vir Bedryfsielkunde, 2003, 29 (2), 39-48

RELIABILITY OF COMPETENCY-BASED, MULTI-DIMENSIONAL, MULTIPLE-RATER PERFORMANCE RATINGS

M DE LANGE
L FOURIE
LJ VAN VUUREN
Programme in Industrial Psychology
Department of Human Resource Management
Rand Afrikaans University

ABSTRACT

The purpose of this study was to test the assumption that the utilisation of competency-based, multi-dimensional, multiple-rater performance ratings would result in reliable and useful measurements of the performance of managers (N=200) in a specific South African organisation. Reliability coefficients were computed and factor analyses were undertaken to determine the reliability of the ratings. The results indicated high inter-method reliability and low inter-rater reliability. Although the competency-based, multi-dimensional, multiple-rater approach appeared to have facilitated valuable input towards the assessment process a greater degree of reliability, validity and usefulness was not necessarily achieved. Implications of these findings are discussed.

OPSOMMING

Die doel van die studie was om die aanname dat die gebruik van bevoegdheidsgebaseerde, multi-dimensionele, meervoudige-beoordelaar prestasiebeoordeling tot betroubare en bruikbare metings van die prestasie van bestuurders (N=200) in 'n spesifieke Suid-Afrikaanse organisasie sou lei, te toets. Betroubaarheidskoëffisiënte is bereken en faktorontledings is uitgevoer om die betroubaarhede te bepaal. Die resultate het hoë inter-metodiese en lae inter-beoordelaarbetroubaarhede aangedui. Hoewel dit wou voorkom of die bevoegdheidsgebaseerde, multi-dimensionele, meervoudige-beoordelaarbenadering tot prestasiebeoordeling nuttige inligting aan die prestasiebeoordelingsproses verskaf, is 'n verhoogde mate van betroubaarheid, geldigheid en bruikbaarheid nie noodwendig verkry nie. Implikasies van die bevindinge word bespreek.

Change within organisations occurs across various dimensions of organisational life. One of the many apparent changes is the gradual flattening of organisational structures from strict hierarchical to less layered or so-called flatter structures (Kanter, 1989). Many researchers predicted that this restructuring of organisations would have definite effects on the functions and performance of managers and professional persons (Cascio, 1995; Greenhaus & Callahan, 1994; Kanter, 1989; May, 1997; Mills, 1991; Naisbitt & Aburdene, 1986; Nolon & Croson, 1995; Pedler, Burgoyne, Boydell & Welshman, 1990). In general, it was argued that managers in these "new" organisations would find it more difficult to supervise their subordinates in the traditional way.

Assessing performance

One of the key functions of supervisory management entails the assessment of past and current performance of subordinates. The appraisal of a subordinate's individual performance in these flattened organisational structures could prove to be especially problematic. Flatter organisational structures would imply fewer managers to appraise the performance of subordinates (Mills, 1991; Naisbitt & Aburdene, 1986; Nolon & Croson, 1995); managers responsible for the appraisal would be responsible for larger groups of workers, and the managers themselves would also have more responsibilities (Hammer & Champy, 1995; McLagan & Nel, 1995). In flatter organisations, workers would have more discretion in how to do their work and clear performance indicators would not exist. The result of this would be that observation and recognition of performance issues, as well as the evaluation of the performance, would become more difficult (Greenhaus & Callahan, 1994). Aspects that would further complicate the function of performance appraisal would be the increased emphasis on team work (as opposed to individual performance), an increase in the number of people working from home in a non-traditional work setting, and various technological developments (Mohrman, Mohrman & Lawler in Bruns, 1992; Nolon & Croson, 1995).

Requests for copies should be addressed to: L. Fourie, Department of Human Resource Management, RAU University, PO BOX 524, Auckland Park, 2006

Despite these obvious challenges to the process of performance appraisal in modern organisations, information generated by performance assessment continues to be vitally important in the people management process within organisations. Performance appraisal information is widely used for decisions on employee counselling, promotions, training, development, salary and bonus allocation, salary administration, personnel audits, identification of potential, job design, work motivation, selection, recruitment, career management and disciplinary actions (Bailey, 1983; Boyatzis, 1982; Cascio, 1991; Goodale, 1993; Mavis, 1994; Philip, 1990; Ricciardi, 1996; Stoner & Freedman, 1989). The search for useful performance appraisal techniques is therefore of an ongoing nature.

Crucial to the search for useful performance appraisal techniques is the quest for reliable and valid measures of performance, yielding information relevant to the decision-making process for which the assessment had been intended. Threats to the reliability, validity, and usefulness of most of the current appraisal techniques are well documented. With specific reference to the performance appraisal of managers, many problems are often encountered (Bailey, 1983; Bocal, 1998; Byars & Rue, 1991; Campbell, 1970; Cascio, 1991; Harvey, 1994; Koontz, 1972; Landy & Fahr, 1983; Levinson, 1990; McLagan, 1994). In this respect it is often found that inter-rater reliability may be low. This may be as a result of influences associated with the method of information gathering, the existence of common rating errors (e.g. halo-effect, leniency, central tendency, recency, bias or subjectivity) and problems related to criterion definition and measurement. In general, validity may be difficult to establish: firstly, because of the difficulties associated with the identification of performance criteria, especially in the case of managers, and secondly, because of the complex nature of managerial behaviour which complicates the process of stating required behaviour in clear behavioural terms. As a result of the fact that managerial performance dimensions are often stated in non-behavioural terms, raters may find it difficult to relate performance to observable behaviours.

However, three relatively recent developments within the domain of performance assessment appear to hold promise for improving the reliability, validity, and usefulness of performance assessments in modern organisations, namely: (1) the use of a competency-based approach to defining desired performance criteria, (2) the use of multi-dimensional assessment procedures, and (3) the use of multiple assessors/raters.

Competency-based approach

The competency-based approach to defining the desired performance criteria in organisations stems from the organisational need to define and develop skills reflecting the current and projected human resource needs of the organisation. It is a critical responsibility of senior management to identify the core competencies of the enterprise, and to ensure that the competencies required by managers, specialists, and the workforce in general, are adequate, appropriate and in line with the mission of the organisation. The development of a comprehensive competency framework and a complementary performance management system is believed to provide an opportunity for enterprise and individual growth, and in the longer term, increased shareholder value.

Advocates for the use of a competency-based approach to performance assessment point to the underlying rationale that a competency can be viewed as a cluster of knowledge, skills, attitudes and behaviours that may be related directly to desired performance on a particular job (Spangenberg, 1990; Spencer & Spencer, 1993; The American Society for Training and Development, 1999). By indicating in clear terms what knowledge, skills, attitudes and behaviours are deemed desirable for optimal performance on a particular job, valid assessment of past and current performance is believed to become possible, and workable indications are supposed to be given of any performance gaps that may exist in terms of the clearly defined competencies/criteria of desired performance.

Spencer and Spencer (1993, p.9) placed special emphasis on the job-relatedness of the competency-based approach by defining competencies as "an underlying characteristic of an individual which is casually related to criterion-referenced effective and/or superior performance in a job situation", while Spangenberg (1990, p.4) added a developmental dimension to the construct of competencies by defining competencies as "the developed abilities, skills, and knowledge managers have acquired through education, training, and experience." It may, therefore, be deduced that the competency-based approach to defining performance criteria for managers, specifically from a developmental point of view, should be considered as a viable option. The competency-based approach is believed to offer a remedy towards the problem of insufficient and non-directive performance assessment data by clearly establishing knowledge, skills, attitudes and behavioural indicators of desired performance.

Apart from the problems encountered in *defining* performance criteria, the actual *assessment* of performance against chosen criteria presents problems of its own. According to Harvey (1994), traditional performance appraisal techniques seldom meet the need for which they were designed; often fail because of their single-dimensional nature; are inherently problematic because they often comprise a single top-down source; provide feedback that are often unclear in developmental terms; and often require skills that raters do not have. Many researchers argued against the use of a top-down approach to performance appraisals. According to Walters (1995), the direct manager is often the least qualified to appraise all aspects of the individual. Williams (in Jones & Bearley, 1996) implied that the restrictions of a top-down approach are real. He argued that no one (rating) source can adequately assess a job holder's performance only because no one source observes all an individual's behaviour. In an attempt to counter some of the problems related to individual assessments, many suggestions implying the use of multi-dimensional multiple-rater ratings have been suggested.

Multi-dimensional, multiple-rater approaches

Leskovec (1967) suggested that a combination of methods be used to assess the performance of managers. Cascio (1991) argued for the use of multi-level, multiple appraiser assessments to appraise the performance of first line managers. Walters and London (in Cascio, 1991) noted that information from multiple sources is more relevant because it will include all the relevant aspects of the job of a first-line manager. This view is supported by Milliman, Zawacki, Norman, Powell and Kirksey (1994). According to these authors, "Multi-level appraisals are becoming imperative in the lean and mean eras where managers have less credibility with their employees due to their larger spans of controls." According to Williams (1989) various sources provide various perspectives on a manager's performance. In response to this, many organisations turn to a multi-level approach (Walters, 1995).

Multi-level appraisal, also known as 360° appraisals, upward appraisal, co-worker feedback, multi-perspective rating and full-circle assessment (Garavan, Morley & Flynn, 1997) can be described as an appraisal process where the participant gets the opportunity to be appraised by various sources. These sources are people with whom the participant has frequent interaction. Appraisal performed by this circle of people is seen as believable, valid, motivational and fair (Edwards, 1998). The appraisers are typically different stakeholders that may include the direct manager, other relevant managers, colleagues, internal and external clients and subordinates of the participant (Jones & Bearley, 1995; Walters, 1995).

It would appear, therefore, that a degree of consensus exists as to the preferred use of more than one appraiser in the assessment of a manager's performance in order to gather reliable, valid and useful information, especially for developmental purposes. Major reasons for including multi-data sources are the cross-validation of perceptions, customer involvement, multi-way management and influence, and the establishment of an improvement agenda (Jones & Bearley, 1995). Nowack (in Garavan et al., 1997) suggested that two of the reasons for the increased use of multi-rater assessments are a need for a cost-effective alternative to assessment centres and, secondly, the need to maximise employee potential in the face of technological change, competitive challenges and increased workforce diversity. The use of multiple raters is also believed to minimise rating biases such as leniency and severity, the halo-effect, failure of discrimination, skewness, extreme-response bias, contrast and similarity effect, as well as the logical error effect (Jones & Bearley, 1995). It is further believed to recognise the complexity of management and the value of input from different sources (Garavan et al., 1997).

Despite the perceived benefits of multiple-rater assessments as described above, Fletcher, Baldry and Cunningham-Snell (1998) argued that "these benefits may be more imagined than real, and that there is no reason to believe that such systems will avoid many of the rating errors and distortions found in traditional top-down appraisal" (p. 19). They further argue that although multiple-rater assessments do negate the subjectivity of ratings, it does not follow that ratings will be accurate, since bias in the form of idiosyncratic rating errors may still be present.

The *assessment centre technique* is essentially a multiple-rater assessment procedure, including the use of multiple sources or appraisers. According to De Beer (1997) assessment centres have been one of the primary areas of development in the field of human resources during the last 30 years. Assessment centres, designed for managerial assessment, assess a person's managerial potential by way of observing his or her behaviour in live managerial situations. The emphasis, therefore, is on actual demonstrated managerial performance and competencies (Spangenberg, 1990). Three elements are central to the assessment centre method. Firstly, the assessment centre is based on the assumption that present behaviour can be used to predict future behaviour. Secondly, it makes use of simulations that are based on the results of precise job analysis. Thirdly, more than one rater is employed in the process

(Jansen & De Jongh, 1997). All the information from the assessment exercises is brought together, and this is usually done under headings of competencies that are perceived to be crucial for high performance in the specific position. These competencies are relevant to the specific position and are based upon information gained from job analysis exercises (Moses & Byham, 1980; Thornton & Byham, 1982; Woodruffe, 1990).

A key feature of assessment centres is the usage of a combination of assessment methods. This implies that the process is rather time-consuming, which makes it costly. Another disadvantage of assessment centres is that they are very labour intensive (Appelbaum, Kay & Shapiro, 1989; Augustyn & Van Wyk,1988; MacDonald, 1988). This feature may, in part, be the reason why multi-rater assessments, such as the 360° degree appraisal technique, are often considered as an alternative.

Extensive research has been done regarding the use of assessment centres, the validity of assessment centres, the cost effectiveness of assessment centres, and the general evaluation of assessment centres (Augustyn & Van Wyk, 1988; Britz, 1984; Charoux, 1991; Dulewicz, 1989; Gaugler, Rosenthal, Thornton & Bentson, 1987; Hunter & Hunter, 1984; Sackett & Ryan, 1985; Spangenberg, Esterhuyse, Visser, Briedenhann & Calitz, 1989; Thornton & Byham, 1982). In general terms, it is believed that assessment centres have acceptable levels of reliability, validity and usefulness, especially in terms of identifying current performance and identifying developmental needs.

To identify the potential and development needs of managers, it is necessary to evaluate managers against clearly established performance criteria. An important requirement for effective performance appraisal is a common understanding of the desired standard of performance (performance criteria) in the job expected from the manager (Philip, 1990). The competency-based approach, inherently part of a typical assessment centre procedure, is therefore often chosen for developmental applications.

A combined approach

Following from the above, it may be hypothesised that the reliability, validity and usefulness of performance assessment procedures could be enhanced by the use of a competency-based approach to defining performance criteria, and by the use of multi-dimensional and multiple-rater techniques such as the assessment centre technique and the 360º degree performance assessment method. Both these procedures represent a multi-dimensional performance assessment by multiple raters, and are therefore believed to have the potential to add value to the assessment process by overcoming at least some of the problems inherent to the more traditional performance assessment procedures.

Given the above assumption, the purpose of the study was to critically evaluate the results of an actual competency-based, multi-dimensional, multiple-rater performance assessment procedure within a South African organisation. The aim of the evaluation was to test the assumption that the utilisation of the above-mentioned approaches/ techniques/methods would result in reliable, valid, and useful measurements of the performance of managers. The first step in the critical evaluation of the actual performance assessment procedure was to investigate the *reliability* of the performance assessment measures. Unless a fair degree of reliability could be proven, questions regarding validity and usefulness would have limited relevance. Only once validity had been established, a critical evaluation of the *usefulness* of the measurement in terms of the intended purpose, namely the identification of developmental needs, could be undertaken.

METHOD

Participants

The research was based on performance assessment data (N=200) gathered within a large South African insurance company. The participants were all existing first-line managers within the company. First-line managers were defined by hierarchical level, with the added requirement of having subordinates reporting to them. The sample included 44% male and 56% female managers, spread across an age range from 23 to 59 years of age.

Procedure

As part of an organisation renewal exercise, desired managerial competencies were previously identified through a process incorporating inputs from human resource specialists, industrial psychologists, senior management consultants and various management focus groups throughout the organisation. Working in a project team and utilising state-of-the-art job analysis technology (the Work Profiling System within the computer-based HR Expert Management Programme), all the inputs were reduced to an organisation-specific competency model for first-line managers. The resultant competency model identified ten competencies believed to represent the domain of desired knowledge, skills, attitudes and behaviours relevant to the successful performance of first-line managers within the organisation (see Table 1).

TABLE 1
ORGANISATION-SPECIFIC COMPETENCY MODEL FOR FIRST-LINE MANAGERS

Competency	Behavioural description of the competency
Developing and empowering others	Actively seeks to improve team member's skills and talents by providing constructive feedback, coaching, training opportunities and assignments that challenge their abilities and encourage development. Delegates responsibilities to appropriate team members, and invests them with the power and authority to accomplish tasks effectively.
Teamwork	Enthuses team members and facilitates successful goal accomplishment by promoting a clear sense of purpose, inspiring a positive attitude to work, sharing information, supporting others and arousing a strong desire to succeed among team members.
Building and maintaining relationships	Able to establish and maintain relationships with people at all levels and from different cultures, puts others at ease; promotes harmony and consensus through diplomatic handling of disagreements and potential conflicts.
Objective setting and management control	Ensures availability of clearly defined objectives and clearly specified action steps for achieving them. Establishes clear priorities; schedules activities to ensure optimum use of time and resources; monitors performance against objectives.
Judgement	Makes rational, realistic and sound decisions based on consideration of all the facts and alternatives available.
Analysis	Seeks all possible information for problem solving and decision making; consults widely, probes the facts, analyses issues from different perspectives. Breaks problems into constituent parts and differentiates key elements from the irrelevant or trivial; makes accurate use of logic, and draws sound inferences from information available.
Commercial orientation	Knowledgeable about financial and commercial matters, focuses on profits, costs, opportunities and activities in order to optimise profitability.
Concern for excellence	Committed to the achievement and maintenance of quality; sets high standards of performance for self and others.
Customer service orientation	Concerned with providing a prompt, efficient and personalised service to clients; goes out of way to ensure that individual customer needs are met.
Decisiveness and execution	Willing to make firm and (if necessary) speedy decisions and committed to definite courses of action; gets results; ensures that key objectives are met.

A comprehensive *competency-based, multi-dimensional, multiple-rater rating procedure* was subsequently developed to evaluate the current performance and/or competence of all the existing first-line managers within the company against the competency model mentioned above. The rating procedure comprised two sets of ratings: a competency-based questionnaire and an assessment centre procedure.

Competency-based questionnaire
The competency-based questionnaire was completed by three different rating sources, namely a *self-rating* by the participant, a rating by the *direct manager* of the participant, and a rating by at least three *subordinates* of the participant. (The average score from the three or more subordinate ratings was used for analysis). The questionnaire comprised clear behavioural descriptions for each of the ten competencies and the rating involved evaluating each participant in terms of observable behaviours usually displayed. Each competency was rated on a five point rating scale with a score of one indicating a significant development need and a score of five indicating outstanding competence. A total score of 50 (with a maximum score of 50 and a minimum score of ten) was computed in respect of each of the three rating sources, producing three independent assessments of each participant for the ten competencies. A *first composite score* with a total score of 150 (with a maximum score of 150 and a minimum score of 30) was obtained by combining the scores from all three the competency-based questionnaire rating sources to present an overall rating for each participant on the ten competencies combined. All the ratings were done independently and anonymously.

Assessment centre procedure
Subsequently, and in addition to the competency-based questionnaire, an assessment centre procedure was developed by professional management consultants to assess the participants across the same ten competencies, in this instance by fully trained assessment centre raters. The assessment centre procedure consisted of three independent exercises, namely: *an in-basket exercise, a role-play exercise, and a structured interview*. Again, a five point rating scale was used with a score of one indicating a significant development need and a score of five indicating outstanding competence. Each in-basket exercise was evaluated by two assessment centre raters, whereas two other assessment centre raters evaluated each role-play exercise and subsequent structured interview. In each instance a score was mutually agreed on by the raters. The assessment centre procedure yielded a score out of five for each competency during each of the three assessment centre exercises. A total score of 50 (with a maximum score of 50 and a minimum score of ten) was computed for each participant in respect of each of the three assessment centre exercises. A *second composite score* with a total of 150 (with a maximum score of 150 and a minimum score of 30) was obtained by combining the scores from the three assessment centre exercises to present a second overall rating for each participant.

Composite and consensus ratings
Two further ratings were computed. In the first instance, the three questionnaire-based ratings were added to the three assessment centre-based ratings to obtain a *third composite rating* with a total score of 300 (maximum rating 300 and minimum rating 60), as an indication of overall competence as assessed by the six independent rating procedures. In the second instance, a moderating committee was established consisting of assessment centre raters, senior managers from the organisation, individual line managers and human resource specialists. All the individual ratings were discussed to agree on a consensus rating out of 5 for each competency in respect of each participant, and a resultant *final consensus rating* with a total score of 50 (maximum rating 50 and a minimum rating 10) for each participant on each of the ten competencies overall.

The competency-based assessment procedure is graphically depicted in Figure 1.

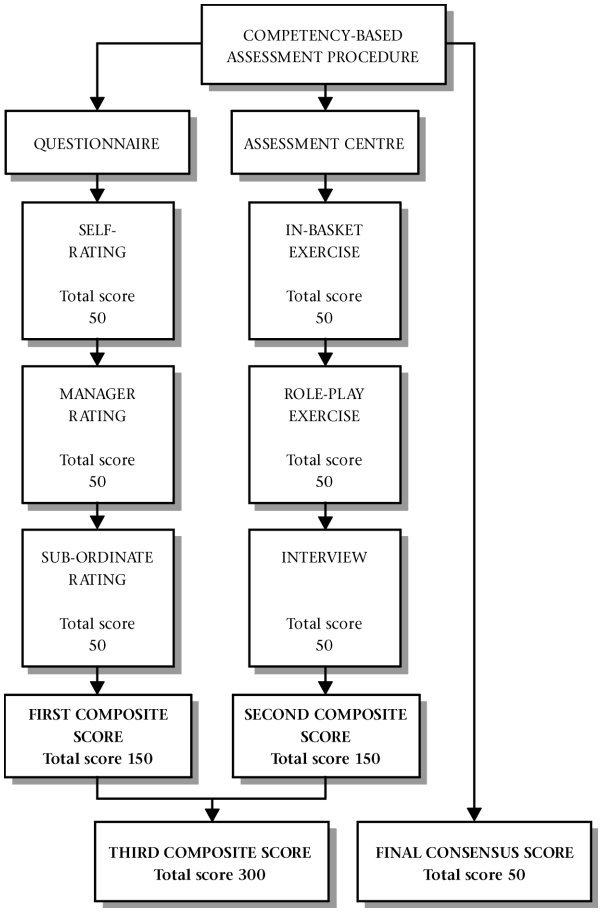

Figure 1: Competency-based, multi-dimensional, multiple-rater assessment procedure

The ten rating scales (i.e. the three questionnaire-based ratings, the three assessment centre-based ratings, the three composite ratings and the final consensus score) obtained for each participant on the ten competencies were viewed as ten different sources of assessment data and these were subsequently subjected to statistical analysis.

Statistical analysis
The data rendered by the rating procedure was analysed by the Statistical Consultation Services of the Rand Afrikaans University using SPSS software. Each of the ten sources of assessment data was critically evaluated in terms of a number of criteria to establish the metrical properties of the different ratings, and the perceived reliability and usefulness of these. The results were inspected to establish possible deficiency with regard to aspects such as failure to discriminate, bias, leniency, strictness and other rating influences.

The statistical analyses included the calculation of average scores, standard deviations, degrees of skewness and kurtosis and ranges of distribution for the different rating scales. Cronbach alpha coefficients were computed to firstly establish the inter-rater reliability among the six different rating sources, and secondly, the internal consistency of ratings within each rating scale. Pearson product-moment correlations were computed to explore the degree and direction of association between the individual rating scales, as well as between the composite rating (obtained by combining the six independent rating scales), and the final consensus rating produced by the monitoring committee. An additional indication of reliability was explored through a principal axis factoring extraction method to uncover a possible emerging factorial structure; firstly, with reference to the ten competencies, and secondly, with reference to the six rating scales. A stepwise regression analysis was subsequently undertaken to establish the contribution of the extracted factors towards the final consensus rating.

RESULTS

The descriptive statistics relating to each of the six independent rating scales, the three composite ratings as well as the final consensus rating are reported in Table 2. These include the mean scores, standard deviations, minimum and maximum scores, ranges of distribution, as well as degrees of skewness and kurtosis in respect of each of the rating scales.

Psychometric properties of the different ratings

With reference to the obtained mean scores, the mean score for the assessment centre procedure overall (Mean = 65,86; SD = 11,07) was substantially lower than the mean score for the competency-based questionnaire overall (Mean = 124,22; SD = 11,22). The self-ratings on the competency-based questionnaire rendered the highest mean score (Mean = 43,39; SD = 4,43), whereas the in-basket ratings of the assessment centre procedure rendered the lowest mean score (Mean = 18,43; SD = 11,07). With reference to the ranges of distribution, the widest range was found in the subordinate and the manager ratings on the competency-based questionnaire. The most restricted range of distribution was observed in the self-ratings on the competency-based questionnaire, as well as the in-basket exercise of the assessment centre. The lowest ratings were attained in the three assessment centre exercises, whereas the highest ratings were attained in respect of the competency-based questionnaires. Inspection of Table 2 indicated various degrees of skewness in the different ratings from the different rating sources. Four of the six rating scales (the three questionnaire-based ratings, as well as the assessment centre interview procedure) produced slightly negatively skewed distributions. Of these, the self-ratings and the subordinate ratings on the competency-based questionnaire were the most skewed. Two of the six rating scales (the role-play and in-basket exercise of the assessment centre procedure) produced distributions slightly positively skewed. The high kurtosis of the subordinate ratings on the competency-based questionnaire was also noticeable.

With reference to the computed *composite ratings* it was found that both the third composite rating (the combined rating from the six independent source) and the final consensus rating revealed enhanced metrical properties, compared to any of the six independent sources of rating. The third composite rating was only slightly negatively skewed and fairly normally distributed. The final consensus rating was even less skewed and more normally distributed. However, ranges of ratings were still fairly restricted.

Inter-rater reliability of the different ratings

The inter-rater reliability of the ratings in respect of the ten competencies is reported in Table 3. Chronbach alpha coefficients were below 0,7 in all instances ((between 0, 24 and 0,54), indicating low levels of agreement between the competency ratings produced by the different rating sources.

TABLE 2
DESCRIPTIVE STATISTICS OF THE TEN RATING SCALES (N=200)

Rating scale	Possible scores between	M	SD	Min	Max	Range	Skewness	Kurtosis
Self-rating	10-50	43,39	4,43	28,70	50,00	21,30	-0,92	0,63
Manager rating	10-50	39,36	5,70	19,90	50,00	30,10	-0,69	0,45
Subordinate rating	10-50	41,47	5,43	18,0	50,00	31,70	-1,23	2,16
In-basket exercise	10-50	18,43	4,28	10,00	30,00	20,00	0,42	-0,16
Role-play exercise	10-50	21,94	5,89	10,00	36,00	26,00	0,18	-0,73
Structured interview	10-50	25,53	4,68	10,00	36,00	26,00	-0,78	0,71
First Composite score (Questionnaires)	30-150	124,22	11,22	81,70	144,80	63,10	-0,88	1,32
Second Composite score (Assessment Centre)	30-150	65,86	11,07	31,00	90,00	59,00	-0,04	-0,26
Third Composite score (Six rating sources)	60-300	190,11	17,18	131,70	230,40	98,70	-0,37	0,36
Final consensus rating	10-50	21,10	3,66	10,00	30,00	20,00	0,13	0,21

TABLE 3
INTER-RATER RELIABILITY OF THE COMPETENCY RATINGS BY THE SIX SOURCES OF RATING

Competency	Roleplay	In-basket	Interview	Selfrating	Manager rating	Subordinate rating	Inter-rater reliability (Alpha)
Developing and empowering others	0,23	0,23	0,32	0,39	0,37	0,33	**0,42**
Teamwork	0,50	0,57	0,44	0,49	0,54	0,43	**0,54**
Building and maintaining relationships	0,29	0,49	0,30	0,41	0,33	0,26	**0,39**
Objective setting and management control	0,46	0,49	0,38	0,42	0,39	0,31	**0,45**
Judgement	0,25	0,40	0,17	0,34	0,31	0,22	**0,33**
Analysis	0,39	0,43	0,29	0,42	0,44	0,32	**0,29**
Commercial orientation	0,51	0,62	0,46	0,54	0,57	0,51	**0,46**
Concern for excellence	0,38	0,44	0,31	0,39	0,36	0,28	**0,41**
Customer service orientation	0,26	0,44	0,20	0,31	0,29	0,27	**0,34**
Decisiveness and execution	0,33	0,52	0,24	0,39	0,37	0,32	**0,24**

Internal consistency reliability
Internal consistency reliability of the competency ratings within each of the six independent sources of rating is reported in Table 4. Chronbach alpha coefficients in this case were all above 0,7 (between 0,75 and 0,97), indicating high levels of agreement between ratings within a particular rating scale. High inter-method reliability is implied.

The low levels of agreement between raters (rating sources) are further apparent from Table 5. Pearson product-moment correlation coefficients indicating the direction and degree of congruence between the ratings produced by the six independent rating sources in respect of each participant are in the very low ranges (r between 0,062 and 0,281). The only exception was found in respect of the correlation between the scores from the structured interview and the role-play (r = 0,691). The ratings produced by the manager and the ratings produced by the role-play showed were negatively correlated; so were the self-ratings and the ratings from the in-basket exercise. The subordinate ratings and the ratings from the in-basket exercise were most highly correlated with the third composite score whereas the ratings from the role-play and the structured interview were most highly correlated with the final consensus rating.

Possible underlying factorial structures
Factor analysis results pertaining to *the ten competency ratings* suggested that a single factor structure emerged within five of the six independent rating sources. Only in respect of the in-basket exercise of the assessment centre did three factors emerge which could not be labelled logically.

Factor analysis results pertaining to the *six sources of rating, however*, attested to the emergence of three factors. The first factor represented the ratings from two of the three assessment centre exercises (the role-play, as well as the structured interview). The second factor represented the ratings from the in-basket exercise of the assessment centre. The third factor represented the ratings from the three competency-based questionnaires. These results implied a high degree of congruence between the ratings of the competencies within exercises, and a far lesser degree of congruence between the ratings within the different sources of rating, once again indicating the possibility of method variance.

A matrix indicating the degrees of congruence between the three extracted factors within the six rating scales on the one hand, and the final consensus score on the other hand, is presented in Table 6. Factor 1 (the role-play and structured interview ratings combined) was most highly correlated with the final consensus rating (r = 0,805). Factor 2 (the ratings from the in-basket exercise), as well as Factor 3 (the ratings from the three competency-based questionnaires combined) showed substantially lower degrees of congruence with the final consensus rating (r = 0, 523 and 0,241 respectively). It is important to note, however, that neither the third composite rating, nor the final consensus rating can be viewed as truly independent criterion measures because both are dependent upon the ratings produced by the assessment procedure itself.

The results of a subsequent stepwise multiple regression analysis is presented in Table 7 indicating that the three factors resulting from the factor analysis together accounted for 84,1% of the variance of the final consensus rating ($F(3,196) = 345.88$, p-value < 0,0005). Factor 1 (the role-play exercise and the structured interview) presented the strongest factor (representing 64,8 % of the variance). Factor 2 (the in-basket exercise) presented the second strongest factor (an additional 18,6 %) and Factor 3 (the three competency-based questionnaires combined) adding only an additional 0,7%.

TABLE 4
INTERNAL CONSISTENCY RELIABILITY OF THE COMPETENCY RATINGS

Competency	Roleplay	In-basket	Interview	Selfrating	Manager rating	Subordinate rating
Developing and empowering others	0,87	0,73	0,86	0,94	0,94	0,97
Teamwork	0,88	0,71	0,86	0,93	0,94	0,97
uilding and maintaining relationships	0,88	0,72	0,87	0,94	0,95	0,97
Objective setting and management control	0,87	0,71	0,87	0,93	0,94	0,96
Judgement	0,89	0,73	0,86	0,94	0,94	0,96
Analysis	0,88	0,74	0,86	0,93	0,95	0,97
Commercial orientation	0,87	0,74	0,87	0,95	0,95	0,97
Concern for excellence	0,87	0,75	0,87	0,93	0,94	0,97
Customer service orientation	0,88	0,74	0,87	0,94	0,95	0,97
Decisiveness and execution	0,88	0,72	0,86	0,93	0,94	0,97
Internal consistency reliability (Alpha)	**0,89**	**0,75**	**0,88**	**0,94**	**0,95**	**0,97**

TABLE 5
CORRELATION MATRIX OF THE SIX INDEPENDENT RATING SCALES

Ratind scales	Self-rating	Manager	Sub-ordinate	Role-play	Interview	In-basket	Third composite score
Self-rating	1,00						
Manager	0,205**	1,00					
Subordinate	0,257**	0,205**	1,00				
Role-play	0,062	-0,019	0,190**	1,00			
Interview	0,141*	0,085	0,281*	0,691**	1,00		
In-basket	-0,061	0,064	0,102	0,128	0,086	1,00	
Third composite score	0,453**	O,527**	0,665**	0,636**	0,688**	0,356**	1,00
Consensus	0,102	0,135	0,271**	0,744**	0,738**	0,523**	0,747**

* Correlation is significant at the 0,05 level (2-tailed)
** Correlation is significant at the 0,01 level (2-tailed)

Table 6
Correlation matrix of the three factors and the composite and consensus ratings

	Factor 1	Factor 2	Factor 3	Composite rating
Factor 1 Pearson correlation Sig. (2-tailed)	1,00			
Factor 2 Pearson correlation Sig. (2-tailed)	0,118	1,00		
Factor 3 Pearson correlation Sig. (2-tailed)	0,178*	0,058	1,00	
Composite rating Pearson correlation Sig. (2-tailed)	0,716**	0,356**	0,770**	1,00
Consensus rating Pearson correlation Sig. (2-tailed)	0,805**	0,523**	0,241**	0.747**

Factor 1: Role-play exercise and Interview
Factor 2: In-basket exercise
Factor 3: Questionnaires
* Correlation is significant at the 0,05 level (2-tailed)
** Correlation is significant at the 0,01 level (2-tailed)

Table 7
Stepwise regression procedure in respect of the consensus rating

Factor	R	R Square	Standard error of measurement
Factor 1 Role-play exercise & Interview	0,805	0,648	2,173
Factor 2 In-basket exercise	0,913	0,834	1,496
Factor 3 Questionnaires	0,917	0,841	1,468

DISCUSSION

A number of meaningful inferences can be drawn from the results. These can broadly be discussed under two headings: the psychometric properties of the different rating scales (sources) and the overall reliability, validity and usefulness of the assessment procedure.

Psychometric properties of the different rating scales (sources)
Scrutiny of the descriptive statistics pertaining to the different rating scales reveals limitations in terms of many of the desired metric criteria. Some scales have relatively restricted ranges and may therefore possibly fail to adequately discriminate among participants (for example: the self-rating and even the final consensus rating). Some scales display high degrees of skewness and kurtosis (for example: the subordinate rating, as well as the role-play exercise) implying once again, not only possible failure to adequately discriminate but possibly also leniency, strictness, or bias.

Differences in the nature of the distributions produced within the six independent rating sources were also noticeable. The difference in the mean scores of the two sets of assessments (the competency-based questionnaires on the one hand, and the assessment centre exercises on the other hand) may serve as an example. The mean score for the questionnaire-based ratings combined is almost double the mean score for the assessment centre exercises combined. A number of reasons may explain this occurrence. Firstly, it might have been possible that the raters of the assessment centre exercises were better trained in the evaluation process and probably more objective. Second, it might have been that the raters in the questionnaire-based ratings knew the participants on a professional and perhaps even on a personal level, predisposing them to a degree of leniency in their ratings. Thirdly, it might have been possible that the assessment centre exercises presented a restricted range of behaviours to assess, limiting the opportunity to credit a participant for displaying particular competencies.

The high mean score of the self-ratings (in comparison to any of the other ratings) serves as another example of the differences in the nature of the distributions produced by the different rating sources. This was to be expected in accordance with previous research findings. It is generally attested that self-ratings are often inflated (Bradley, 1978; Fox, Caspy & Reisler, 1994; Snyder, Stephan & Rosenfield, 1976). A study by Theron and Roodt (1999) specifically indicated that self-ratings are inflated, unreliable, invalid, biased and generally suspect when compared to the ratings of other raters. This apparent leniency may be attributed to defensiveness and a desire to enhance perceptions of the self (Holzbach, 1978; Steel & Ovalle, 1984). Thornton (1980) indicated that self-ratings should therefore be used cautiously.

The lower mean score of manager ratings, on the other hand, might possibly be explained by managers having a higher status than their subordinates, thus expecting their subordinates to conform to their own standards (Harris & Schabroeck, 1988). It might also have been that different perceptions of the role requirements for managers existed among raters at different levels in the organisational hierarchy (Theron & Roodt, 1999). Raters at different levels may emphasise different dimensions of performance and arrive at differential assessments (Landy, Farr, Saal & Freytag, 1976; McEnergy & Blanchard, 1999). In addition, individuals might have felt threatened because the assessment procedure was new to everyone, and subordinates might have felt threatened that their managers could retaliate if the upward feedback results were not favourable (London, Wohlers & Gallagher, 1990). This may also explain the unacceptably high kurtosis of the subordinate ratings implying that this rating scale in particular probably failed to adequately and meaningfully discriminate between the participants in terms of individual competence. A mentioned earlier, the distribution was also distinctly skew, implying that very few participants were rated negatively. A more platycurtic distribution would clearly have been more desirable.

With reference to the general metrical properties of the different rating scales, therefore, it may be argued that none of the independent rating scales presented distributions that could be viewed as ideal. Both the third composite rating (combining the six independent ratings) and the final consensus rating appear to possess somewhat enhanced psychometric properties: acceptable degrees of skewness and kurtosis and fairly acceptable ranges of distribution. This could be interpreted to imply that these rating sources possibly render measures more acceptable in terms of their *psychometric properties* in comparison to any of the six independent rating scales. It could also be argued that many of the common rating errors of individual ratings appear to have been countered by the process of "adding together" input from different rating sources.

However, the low levels of agreement that generally prevailed amongst the ratings from the different rating sources (low inter-rater reliability) places a question mark over the reliability, validity, and usefulness of any of these measures, regardless of their apparent enhanced metric properties.

Reliability, validity, and usefulness of the ratings
The fact that such low levels of congruence were found between the six independent sources of rating is of grave concern. Whilst reasons can be found for the differences found in the ratings from the different rating sources, it does not bode well for the overall reliability, validity, and usefulness of the assessment procedure.

It might, for instance, be possible to explain the low levels of agreement between the competency-based questionnaire ratings and the assessment centre exercise ratings by suggesting that raters were informed by different behaviours from the participants, or that questionnaire ratings were based on behaviours in the past, relying on memory that could have presented an incomplete recollection of past behaviour; assessment centre exercise ratings, on the other hand, were based on observable behaviours as they occurred during the assessment centre simulations. It might further be argued that raters do not necessarily share the same meaning attached to the different competencies or terminologies used in the descriptions of the different behaviours (London & Smither, 1995; Theron & Roodt, 2000). Furthermore, raters might have been unable to interpret certain behaviours according to the defined behavioural descriptions.

With reference to assessment centre procedures specifically, methodological problems regarding construct and content validity are often found (Klimoski & Brickner, 1987; Sackett & Dreher, 1982; Sackett & Hakel, 1979). In the current study it might have been possible that the sample of behaviour evaluated during the assessment centre exercises was not sufficiently representative to base ratings on – especially with the aim of identifying development needs. A further complication might have arisen as a result of the fact that the ratings for the structured interviews and the ratings for the role-play exercise were done by the same raters, allowing for a degree of bias manifesting in the ratings. This might explain the highly correlated ratings from these two rating sources. It might also partly explain the apparent weight that these two ratings carried in terms of the final consensus score; if these two assessment centre raters had the opportunity to convincingly present their combined views on a participant to the final rating committee, further bias might have been introduced to the rating process. It is difficult to imagine that ratings produced by sources familiar with the actual (versus simulated) behaviours of the participants – managers and subordinates – failed to produce meaningful and valid inputs into the assessment process, and in carried such limited weight in terms of the ratings of the final consensus score.

It is clear, therefore, that the low inter-rater reliabilities between the different sources of rating could have been influenced by at least four sources of error variance: content sampling, heterogeneity of the behaviour domain sampled, rater bias and method variance. In a multi-trait, multi-method study conducted by Shore, Shore and Thornton (1992), method variance was also shown to be problematic: competency ratings within a particular rating scale tended to be more highly correlated than were the ratings of the same competency across rating scales.

A further reason for the low inter-rater reliabilities might have been the nature of the rating scale used. In the current study, a five-point scale was used. Downie and Heath (1976) were of the opinion that, in order to yield maximum reliably ratings, it is generally preferable to start out with as refined a category system as raters are capable of discriminating, and then to later collapse adjacent categories, if found necessary. It might have been worthwhile to use a seven-, or nine-point rating scale of intensity.

The high internal consistency reliabilities found in the current study are far from ideal. These high correlations suggest that the competencies may not be independent from one another and that they are all highly inter-correlated; alternatively, that raters were unable to differentiate between different dimensions of managerial behaviour. This places a question mark on whether the descriptions of the competencies were done comprehensively enough to distinguish between the behaviours observed. It is also realistic to assume that halo-effects, as a result of many inter-relationships among competencies, may have complicated the assessment of distinctly separate competency scales. For example: a good communicator may have been judged to also be a good team member, simply because communication ability is critical to being a good team member.

The results of the principal axis factor analysis performed on the ten competencies which produced only one factor, may confirm the possibility that the assessment either dealt with a single-dimensional construct (managerial competence) or that raters had difficulty distinguishing the distinct competencies from one another. This concurs with the widely held notion that management is indeed a single, yet multi-faceted construct. Stoner and Freeman (1989) suggested that management is a single dimensional process because all managers, regardless of their particular aptitude or skills, engage in certain inter-related activities in order to achieve their desired goal. Cascio (1991) further argued that a single factor emerges as a result of the assumption that a general factor within all of the criteria accounts for virtually all of the important variance in management behaviour.

However, the subsequent factor analysis computed in respect of the six independent sources of rating (resulting in three factors), clearly suggested that the rating procedure/method might have been very important in explaining the variances found. The different methodologies used in each of the rating scales: namely questionnaires, interactive approaches (the interview and role-play exercise) and a simulated written approach in the in-basket exercise, could have influenced the ratings in many ways. It is evident that the raters of the in-basket exercise, for instance, had unique perspectives on the behaviours that were important, and related to the desired competencies described in behavioural terms. A study done by McEnergy and Blanchard (1999) suggested that raters from assessment centre exercises often had limited exposure to participants and because of this, may have been less able to put the participants' behaviour into the appropriate context. It might also have been that raters measured behaviour differently and under different circumstances (Scullen, Mount & Goff, 2000; Viswesvaran, Ones, & Schmidt, 1996). Steele and Ovalle (1984) stated that inconsistencies between ratings could only be minimised if the evaluative criteria are clearly defined, thus establishing a common frame of reference in the rating procedure.

What is of concern, however, is that the assessment procedure in its entirety, failed to unambiguously identify a particular training need. This may be illustrated by an example of the actual ratings of a particular participant across all six independent rating scales. This participant attained a score of one for the teamwork competency in both the role-play and the in-basket exercise of the assessment centre procedure; a score of two in the structured interview; and score of four and above in each of the three competency-based questionnaire rating scales. The incongruence of these scores for a particular participant on the teamwork competency is clear from Table 8 below.

Table 8
Ratings of the teamwork competency for a single participant

RATING SOURCE	Min score	Max score	Actual score	M
Role-play	1	5	1	
Structured interview	1	5	2	
In-basket	1	5	1	
Self-rating	1	5	5	
Subordinate rating	1	5	4	
Manager rating	1	5	4	
FINAL CONSENSUS RATING	1	5	2	
First composite rating (Questionnaires combined)	3	15	13	4,3
Second composite rating (Assessment centre combined)	3	15	4	1,3
Third composite rating (Six ratings combined)	6	30	17	2,8

The question can rightfully be asked whether the above implies a development need, or not. Note that the consensus score for this participant was two (implying a clear development need), despite the fact that both his manager, as well as his subordinates, indicated otherwise. If inter-rater reliabilities are indeed this low, validity is obviously at stake and practical usefulness is resultantly limited.

Implications

It is important to understand that opportunities to succeed or fail in implementing a multiple-rater performance assessment system may occur at every stage of the process, from the design and planning phases, to the development of the instrument, the instrument design, the administration, the feedback processing and reporting, as well as the overall action planning (Bracken, 1994). Despite the concerns raised above, the results of the stepwise multiple regression analysis pertaining to the final consensus rating, appear to indicate that the competency-based, multi-dimensional, multiple-rater approach to performance assessment nevertheless added value to the assessment procedure. It is clear that the different sources of rating all added value to the process in terms of their seemingly different perspectives when evaluating the behaviour of the participants. Whilst by no means a flawless procedure, the study confirms the theory that data from multiple sources are desirable because it provides a more complete picture of the individual's strengths and weaknesses, focusing on different aspects of performance as acknowledged by different raters (Cascio, 1991; Jones & Bearly, 1996). The high correlation between the third composite rating and the consensus rating (r = 0,747 significant at the 0,01 level) serve as a further confirmation that multiple views of a person's behaviour may compensate for many of the apparent shortcomings of any single-rater assessment procedure.

CONCLUSION

Primarily two approaches to improve performance appraisals ratings overall, have been suggested by various researchers, namely rater training and scale development along the lines of psychometric requirements (Fletcher et al., 1998; Lievens, 1998; Woehr & Huffcutt, 1994). These authors indicate that the nature of the rating instruments used may have an effect on the cognitive processes involved, such as observing, storage and retrieval of information. According to Kriek (1991) and McEnergy and Blanchard (1999) a common mental model of a particular level of performance and how important different dimensions of performance are, should be clarified and emphasised, if multi-rater assessments are to be useful for development. It is also important to understand the possibility of different perspectives prevailing among raters (Theron & Roodt, 2000). According to Tornow (1993) it may be less a question about who are right and more a question of what various perspectives can contribute to the understanding of an individual's strengths and weaknesses. It appears that multi-rater assessments may fulfil the need of providing individuals with more holistic information about their performance, in order to facilitate development.

The study confirms the notion that the reliability, validity and usefulness of performance assessment procedures may be enhanced by the use of a competency-based approach to defining performance criteria, and by the use of multi-dimensional and multiple-rater techniques, such as the assessment centre procedure and the 360º or multiple-rater performance assessment method. Whilst the findings of the study may have been compromised by features of the research design, such as the use of different managers to rate the different participants, the inferences appear to be credible enough, to warrant further research towards establishing the overall validity of these assessments for the purposes intended, possibly through a well-designed criterion referenced study.

ACKNOWLEDGEMENT

The authors would like to thank Riëtte Eiselen and James du Toit of Statcon at RAU, for their professional service and valuable contribution regarding the data analysis of this project.

REFERENCES

Appelbaum, S.H., Kay, F. & Shapiro, B.T. (1989). The assessment centre is not dead! How to keep it alive and well. *Journal of Management Development*, *8*, 51-56.

American Society for Training and Development (ASTD). (1999). www.pbs.infopro.net" *www.pbs.infopro.net*.

Augustyn, J.C. & Van Wyk, A.J. (1988). Die Sestien Persoonlikheidsfaktorvraelys (16PF) as hulpmiddel by die takseersentrum. *Journal of Industrial Psychology*, *14*, 25-27.

Bailey, J. (1983). *Job design and work organization*. Englewood Cliffs, NJ: Prentice-Hall.

Bacal, R. (1998). *Why employee ranking systems lead to disaster*. Retrieved from Infoseek database on the www.escape.ca/rbacal/articles.htm

Boyatzis, R.E. (1982). *The competent manager*. New York: Wiley.

Bracken, D.W. (1994). Straight talk about multi-rater 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.

Britz, P.J. (1984). *Die validering van 'n bestuursbeoordelings-sentrum*. Unpublished doctoral dissertation, University of Pretoria, Pretoria.

Bruns, W.J. (1992). *Performance measurement, evaluation and incentives*. Boston: Harvard Business School.

Byars, L.L. & Rue, L.W. (1991). *Human resource management*. Boston: Irwin.

Campbell, J.P., Dunnette, M.D., Lawler, E.E. & Weick, K.E. (1970). *Managerial behavior, performance and effectiveness*. New York: McGraw-Hill.

Cascio, W.F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cascio, W. F. (1995). Whither industrial/organizational psychology in a changing world of work? *American Psychologist*, *50*, 928-939.

Charoux, E. (1991). Assessment centres: A partial objective look. *Human Resource Management*, *7*, 28-29.

De Beer, N. & Van Vuuren, L.J. (1997). Selfbeoordeling as 'n voorspeller van waargenome gesimuleerde bestuursvermoë, soos gemeet in 'n takseersentrum. *Journal of Industrial Psychology*, *23*, 12-20.

Downie, N.M. & Heath, R.W. (1976). *Basic statistical methods*. New York: Harper and Row.

Dulewicz, V. (1989). Assessment centres as the route to competence. *Personnel Management*, *21*, 56-59.

Edwards, M.R. (1998). Improving performance with 360-degree feedback. *Career Development International*, *1*, 13-25.

Fletcher, C., Baldry, C. & Cunningham-Snell, N. (1998). The psychometric properties of 360 degree feedback: An empirical study and a cautionary tale. *International Journal of Selection and Assessment*, *6* (1), 19-34.

Fox, S., Caspy, T. & Reisler, A. (1994). Variables affecting leniency, halo and validity of self-appraisal. *Journal of Occupational and Organizational Psychology*, *67*, 45-46.

Garavan, T.N., Morley, M. & Flynn, M. (1997). 360 degree feedback: its role in employee development. *Journal of Management Development*, *16*, 134-147.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C. & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, *72*, 493-511.

Goodale, J.G. (1993). Seven ways to improve performance appraisals. *HR Magazine*, May, 77-80.

Greenhaus, J.H. & Callanan, G.A. (1994). *Career management*. London: Dryden.

Hammer, M. & Champy, J. (1994). *Reengineering the corporation: A manifest for business revolution*. New York: Brealey.

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

Harvey, E.L. (1994). Turning performance appraisals upside down. *Human Resource Professional*, *7*, 30-32.

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

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72-92.

Jansen, P. & De Jongh, F. (1997). *Assessment centers – A practical handbook*. New York: Wiley.

Jones, J.E. & Bearley, W.L. (1996). *360 degree feedback*. Amherst: HRD Press.

Kanter, R.M. (1989). The new managerial work. *Harvard Business Review*, November-December, 85-92.

Klimoski, R. & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology*, *40*, 243-260.

Koontz, H. (1972). Short-comings and pitfalls in managing objectives. *Management by Objectives*, *1*, 6-12.

Kriek, H.J. (1991). Die bruikbaarheid van die takseersentrum: 'n Oorsig van resente literatuur. *Journal of Industrial Psychology*, *17*, 34-37.

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

Landy, F.J., Farr, J.L., Saal, F.E. & Freytag, W.R. (1976). Behaviorally anchored scales for rating the performance of police officers. *Journal of Applied Psychology*, *61*, 750-758.

Leskovec, E.W. (1967). A guide for discussing performance appraisal. *Personnel Journal*, *46*, 150-152.

Lievens, F. (1998). Factors which improve the construct validity of assessment centres, A review. *International Journal of Selection and Assessment*, *6*, 141-152.

London M. & Smither, J.W. (1995). Can multi-source feedback change perceptions of goals accomplishment, self-evaluations, and performance-related outcomes? Theory-based applications and directions for research. *Personnel Psychology*, *48*, 803-839.

London, M., Wohlers, A.J. & Gallager, P. (1990). 360 degree feedback surveys: A source of feedback to guide management development. *Journal of Management Development*, *9*, 17-31.

MacDonald, D.R. (1988). Greater results from your assessment center. *Training and Development Journal*, *42*, 50-51.

Mavis, M. (1994). Painless performance evaluations. *Training and Development Journal*, October, 40-44.

May, K.E. (1997). *Work in the 21st Century: Implications for performance management*. Retrieved from the Society of Industrial and Organizational Psychology database on www.siop.org

McEnergy, J.M. & Blanchard, P.N. (1999). Validity of multiple ratings of business student performance in a management simulation. *Human Resource Development Quarterly*, *10*, 155-172.

McLagan, P. (1994). Performance management: Can it work? *HR Magazine*, September, 23-25.

McLagan, P. & Nel, C. (1995). *The age of participation*. Randburg: Knowledge Resources.

Milliman, J.F., Zawacki, R.A., Norman, C., Powell, L. & Kirksey, J. (1994). Companies evaluate employees from all perspectives. *Personnel Journal*, November, 99-103.

Mills, D.Q. (1991). *Rebirth of the corporation*. New York: Wiley.

Mohrman, A.M., Mohrman, S.A. & Lawler, E.E. (1992). The performance of teams In W.J. Bruns (Ed.). *Performance measurement, evaluation and incentives*. Boston: Harvard Business School Press.

Moses, J.L. & Byham, W.C. (1980). *Applying the assessment center method*. New York: Pergamon.

Naisbitt, J. & Aburdene, P. (1986). *Re-inventing the corporation*. New York: Warner.

Nolon, R.L. & Croson, D.C. (1995). *Creative destruction*. Boston: Harvard Business School Press.

Pedler, M., Burgoyne, J., Boydell, T. & Welshman, G. (1990). *Self-development in organizations*. New York: McGraw-Hill.

Philip, T. (1990). *Appraising performance for results*. London: McGraw-Hill.

Ricciardi, P. (1996). Simplify your performance measurement. *HR Magazine*, March, 98-106.

Sackett, P.R. & Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology*, *67*, 401-410.

Sackett, P.R. & Hakel, M.D. (1979). Temporal stability and individual differences in using assessment information to form overall ratings. *Organizational Behavior and Human Performance*, *23*, 120-137.

Sackett, P.R. & Ryan, A.M. (1985). A review of recent assessment centre research. *Assessment Centers and Management Development*, *4*, 13-27.

Scullen, S.E., Mount, M.K. & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*, 956-970.

Shore, T.H., Shore, L.M. & Thornton, G.C. (1992). Construct validity of self and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology*, *77*, 42-54.

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

Spangenberg, H.H. (1990). *Assessing managerial competence*. Cape Town: Juta.

Spangenberg, H.H., Esterhuyse, J.J., Visser, J.H., Briedenhahn, J.E. & Calitz, C. (1989). Validation of an assessment centre against BARS – an experience with performance related criteria. *Journal of Industrial Psychology*, *15*, 1-10.

Spencer, L.M. & Spencer, S.M. (1993). *Competence at work*. New York: Wiley.

Steele, R.P. & Ovalle, N.K. (1984). Self-appraisal based upon supervisory feedback. *Personnel Psychology*, *33*, 263-271.

Stoner, J.A. & Freedman, R.E. (1989). *Management*. London: Prentice-Hall.

Theron, D. & Roodt G. (1999). Variability in multi-rater competency assessments. *Journal of Industrial Psychology*, *25* (2), 21-27.

Theron, D. & Roodt G. (2000). Mental models as moderating variable in 360 degree competency assessments. *Journal of Industrial Psychology*, *26* (2), 21-27.

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

Thornton, G.C. & Byham, W.C. (1982). Assessment centers and managerial performance. New York: Academic Press.

Tornow, W.T. (1993). Editor's note: Introduction to special issue on 360-degree feedback. *Human Resource Management*, *32*, 211-230.

Viswesvaran, C., Ones, D.S. & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557-574.

Walters, M. (1995). *The performance management handbook*. London: Institute of Personnel and Development.

Woehr, D.J. & Huffcutt, A.J. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organisational Psychology*, *67*, 189-205.

Woodruffe, C. (1990). *Assessment centres*. London: Institute of Personnel Management.