MULTI-SOURCE PERFORMANCE APPRAISAL: AN EMPIRICAL AND METHODOLOGICAL NOTE

YEHUDA BARUCH
The Technion, Israel University of Technology
GEDALIAHU HAREL
University of Calgary, Canada

INTRODUCTION

Performance appraisal is an important issue in the area of Human Resource Management (HRM). There is considerable agreement as to the centrality and importance of the performance appraisal process among scholars concerning both public (Heneman et al., 1989; French, 1987) as well as private organizations (Foltz et al., 1983; Lacho et al., 1991). We may learn of the importance attached to the performance appraisal system through the large number of organizations that use various kinds of formal performance appraisal procedures—about 90% of all organizations (Lazer and Wikstrom, 1977; Varney, 1972). There is a consensus that performance appraisal is a key process which is vital to the existence and development of the organization (Latham and Wexley, 1981; Fombrun and Lau, 1983).

The term assessment is associated with a variety of concepts, programs, and procedures. It includes measurement, evaluation, correlation, and prediction with one or more variables and criteria, as well as standard of performance (Herrold, 1986). Illustratively, Milkovich and Boudreau (1988) defined performance appraisal as the process that measures employee performance, which involves four decisions: (1) what to assess; (2) who should make the assessment; (3) which assessment procedures to use; and (4) how to communicate assessment results.

This article focuses on the third decision and presents an innovative approach that combines information from several sources in order to reach more valid and reliable performance appraisal (Latham and Wexley, 1981; Schneier, 1977; Tsui, 1984; Tsui and Ohlott, 1988). This approach may contribute to a reduction in performance appraisal problems which are usually due to poor information from one or more of the performance appraisal sources. The authors will not consider other appraisal problems such as inadequate descriptions of criteria or difficulties in methods.

There are several objectives in employing a performance appraisal system. The main one is to satisfy the need of the organization for information about its employees’ abilities, qualifications, training needs, and potentials. Another objective is to provide feedback to the workers. In order to serve all of these needs, the performance appraisal system must be valid and reliable.

There are several ways to improve the performance appraisal system. The common recommendation for improvement of performance appraisal systems is to base them on behavioral scales instead on traits evaluation (Murphy and Cleveland, 1991; Heneman et al., 1984). Another recommended method is to base the appraisals on more than one source of information. There are, however, difficult problems in developing such a composite model and most organizations use only one source of information—the direct supervisor, as reported in Field and Holley (1977) and in Lazar and Wikstrom (1977). This problem limits the feasibility of implementing and evaluating the effectiveness of a composite criteria model.

Literature review uncovered several sources of information available for appraising public sector employees, with the most common ones presented in Figure 1 (Ammons, 1987). The main one is the direct manager’s appraisal of his or her subordinates. Other possible sources are subordinate’s appraisal or self-appraisal (Steele, 1985), objective information, and sources outside the organization (such as independent advisors or customers, suppliers, etc.). All relevant sources have shortcomings affecting the validity and reliability (see Bernardin, 1986; Ashford, 1989). In addition, it is difficult to ascertain the relative validity and reliability of the sources.

Some researchers have recommended that the quality of performance appraisal could be improved by combining sources but none suggest a method for accomplishing this task (Schneider and Carrol, 1977; Miner and Miner, 1985; Kavanagh, 1982). Borman (1980), as cited in Kavanagh (1982) argued that, although it is possible to gain more information, accuracy may not increase when multiple sources are used. This claim may be correct in the cases where the correlation among raters is low. One should not leap to
the conclusion that this kind of situation (low correlation among raters) indicates low validity. In contrast, different rating sources have different perspectives and different opportunities to observe rated behavior. Some have argued that this may result in relatively low inter-rater agreement but with each rating source having a high degree of validity.

In fact, when examining research on rating sources, the authors find that different sources often give similar ratings. In a meta-analysis of research on rating sources, Harris and Schaubroeck (1988) found quite high correlations (.62) between peer and supervisory ratings. In addition, Weckley and Gier's (1989) work indicates that there might be a ceiling for the level of accuracy for the different raters. One might expect higher accuracy in the case where the raters are experts either in performing or supervising the performance of the job in question or both. The experts should have experience and training as raters in this area and must have the ability to observe the subjects.

All the above conditions were fully present in the research by Weckley and Gier (1989) which was based on rating sportsmen by expert raters in the Olympic games. Even in this case, they found the inter-rater reliability of two judges to be between .76-.88 (average).

In the case where different raters have different points of view and different perceptions and attitudes, we may expect, at most, moderate correlations (Tsui and Ohlott, 1988). Such a level of correlation may indicate that the different raters are measuring the same dimension of performance. Thus, it is reasonable to assume that the achievement of moderate correlations enables us to improve the accuracy of the performance appraisal process by combining multiple raters' appraisals into one performance appraisal (see also in Edward and Sproull, 1985).

The seminal article on composite versus multiple criteria (Schmidt and Kaplan, 1971) reviewed a multitude of research on the long-standing controversy in industrial psychology between the advocates of the single composite criterion and those favoring multiple criteria. After considering the argument on both sides, Schmidt and Kaplan (1971:431) propose that both approaches are legitimate for their own purposes and conclude:

Criterion elements can be, and, in fact at some point must be, weighted into a composite irrespective of their intercorrelations. If all criterion elements are considered to be measures of a single underlying economic
construct, the resulting composite unambiguously represents the economic construct and is interpreted as such.

In a follow-up study, James (1973) suggested that, in order to develop a general criterion model, "various sources of measurement should be employed in obtaining criterion data." He adds (Ibid., 78):

These sources could include behaviorally-based job performance ratings from different raters; objective measures of job performance, ability, motivation, and situational data; and global measures of performance.

Developing a Multiple-Sources Model

In this article, the authors will develop a general model of combining multiple-source performance appraisal which should improve the quality and validity of the performance appraisal process. Following a discussion of the model they will present a case study in which the model is demonstrated.

The construction of the model includes the following seven main steps:

1. Identifying all feasible and existing sources that may be used in the process of performance appraisal for the assessed population;
2. Determining and choosing those appraisal sources which are relevant to the performance dimension that the performance appraisal system aims to assess;
3. Gathering samples of appraisals from these sources and examining the correlations between them;
4. Testing for high correlations and deleting the sources that do not contribute more than others; i.e., if there is a very high correlation between two sources, it is possible to use the average or to omit one of them, especially if collection of the information is costly;
5. Testing for low correlations. If the correlations between one source and the others are found to be very low, there are two possible explanations. One possibility is that the low correlations are due to lack of reliability and/or validity of one of the sources, in which case the information from that source should be ignored. The other possibility is that the source contains information on performance dimension that other sources do not include or emphasize, in which case this source should be considered as a distinct source and should be included in the analysis;
6. Collection of the appraisals of all remaining sources for every employee in the appraised population; and
7. Combining the different appraisals into one total score by multiplying each source score by a weighting factor.

It is possible to define several scores relating to different performance dimensions. All of the seven steps should be carried out for each score.

The evaluation factors should be defined differently for every sector of the population, if necessary (e.g., managers, unskilled workers, etc.) and differently for every performance dimension. In order to define the factors, the scores of all the different sources should be standardized to the same scale, unless all are on the same metric. The recommended scale is one based on standard scores or simply arraying all of the sources on a scale of 1-100.

There are several approaches for defining the relative weight for each source. It is possible to use a simple additive measure, although this is an uncommon occurrence in most organizations. The second possibility, if there are enough sources which is seldom the case, is to use a factor analysis. A third possibility is to use managers from within the organization and ask for their opinions about the preferred relative weight. Thus, a forum of "decision-makers" in the organization seems to be the most appropriate and convenient group to use for deciding on the weighting formula. Finally, the definition of the relative weights can be done by external experts.

It would be difficult to determine which of these four approaches is superior or more efficient, absent an objective measure. However, such "objective" measures of performance are hard to come by.

In the following section, the authors will describe the use of the above model in the performance appraisal process of technicians in the Israeli Air Force (see also Harel and Baruch, 1991).

SUBJECTS

The participants were two groups of technicians who had graduated from two different five-year training programs and then joined the Israeli Air Force (IAF) in technical positions. Members of one group were trained in the IAF vocational school (about 200 men). In
this school, students studied technical subjects as well as military issues related to the IAF. The students could choose to stay in the boarding school, which most of them did. The environment of the boarding school fosters an improved atmosphere of learning and attachment to the IAF culture. The second group included individuals from a representative sample trained in other civilian schools in Israel (about 500 men).

The background characteristics of the members of both groups are similar—e.g., age, years of education, intelligence, and motivation. This similarity was verified through the use of a special evaluation scale that is administered to every soldier in the Israeli Defense Forces, including the IAF. The data for this score consists of a weighted formula that combines information on personal socioeconomic status, education, intelligence, motivation, and language ability. The scores of this evaluation scale were compared for both populations and no significant differences were found.

All information was gathered from the personal files of the participants and the authors had permission to access any data relevant to the research. The information about each participant was gathered from the time of conscription to the IAF at the age of 18 through the fourth year of services.

MEASURES AND PROCEDURES

The authors used a special procedure for combining the separate information from the various appraisal sources into one score. In order to determine the weight for the various sources, they asked a sample of 50 high-ranking officers (colonel and higher) in the IAF to weigh the various sources of performance appraisal. They asked the officers what should be the suitable weight for each score. The results obtained from this procedure determined the relative weight of the various sources of the evaluation used in this study.

The final weights were:

For the technical performance evaluation: 35%-direct supervisor evaluation; 35%-peer evaluation, and 30%-course grades. For overall performance: 60%-direct supervisor evaluation, and 40%-peer evaluation.

The statistical analysis consisted of three stages. The first stage computed a correlation matrix for the different sources of appraisal on both dimensions—technical and overall performance. The second stage computed scores to obtain one final score in every dimension for each participant in the study. The last stage consisted of a series of T-tests comparing the averages of the two populations as well as of F-tests comparing the variances. The outputs were the average and the standard deviation of the dependent variables—performance appraisal, professionalism, and organizational and professional commitment.

RESULTS

The model produced scores on two performance dimensions: technical performance and overall performance. In order to determine the technical performance dimension, the authors used the following procedure: in the direct supervisor’s performance appraisal form there are two performance scales in which the evaluator is asked to rank the professional knowledge and professional ability of the employee (i.e., the IAF soldier). In the peer’s sociometric form there is a separate clause for the evaluation of professional level. In addition to these sources, there are the individual performances in all the technical courses taken during the service of each technician in the IAF. The grades in these courses are kept in a computerized database.

In order to determine the overall performance dimension, the authors used the following procedure. In the direct supervisor’s performance appraisal form there are several clauses in which the evaluator is asked for his or her opinion related to the general functioning of the soldier and, in addition, a summarized general score for overall performance as a soldier (employee). An equivalent clause appears in the peer’s sociometric form to be filled out by all the peers in the working team of every soldier.

Score Correlations

The correlation between the various sources was found to be relatively high, especially between the direct supervisors’ and the peer appraisals and less so between the course grades and the two former sources. A possible explanation is that the latter refers more to theoretical knowledge and the former to actual work. This explanation applies to both military and civilian sub-populations.

Tables 1 and 2 present the correlations between the different sources of information on two dimensions—technical performance and overall performance for the entire population. Table 1 presents
the correlations between the three sources of appraisals and between these sources and the final computed technical performance score.

Some of the reasons for the moderate correlation coefficient between the direct supervisor's performance appraisal and the peers' performance appraisal is the ambiguity of the definitions and different points of view of different sources. As noted earlier, the authors could not expect very high correlations from these sources. They, therefore, used the computed scores as a basis for comparison between the two groups. The results in the last column of Tables 1 and 2 enable them to combine the scores to get a grand total score.

In order to show the advantages of the use of the model, the authors present Tables 3 and 4, simple comparisons for each score. Then, using the weights mentioned earlier, they compared the final scores for the whole population. The results of this comparison are presented in Table 5.

Tables 3 and 4 show no significant differences in the performance appraisal of the IAF graduates as compared with the civilian school graduates except for course grades. Here the IAF graduates scored significantly higher than the civilian school graduates. Combining the various sources of performance appraisal into one single scale, as the model suggests, enables the authors to reach decisive conclusions as to the differences between the two groups.

The results in Table 5 show significant differences between the two groups of subjects. IAF graduates were found to be better in technical performance ($\bar{X} = 85.2$ vs. $60.2$ on 0-100 scale) and also better in overall performance ($\bar{X} = 57.6$ vs. $54.2$). Assuming that one possible justification for having a specialized school is getting better-equipped employees, identification and development of employees of high potential for the organization is considered to be of great importance to high technology organizations in particular (Derr et al., 1988).

From the difference in the standard deviations, the variability within the group of IAF graduates is smaller than within the group of the civilian school graduates. It means that the workforce trained in the IAF school shows more consistency in quality of performance than those trained in the civilian schools, thus able to enhance the managers' ability to predict performance.

To summarize the case, the authors found significant and meaningful differences between the two populations when combining technical and overall performance. These differences were not clearly evident when they used the separate sources.
TABLE 4
T-TEST OF DIFFERENT OVERALL PERFORMANCE SCORES OF
IAF TECHNICAL SCHOOL GRADUATES AND CIVILIAN
TECHNICAL SCHOOL GRADUATES

<table>
<thead>
<tr>
<th></th>
<th>IAF School Graduates</th>
<th>Civilian Tech. School Graduates</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Supervisor Score</td>
<td>85.3</td>
<td>85.3</td>
<td>n.s.</td>
</tr>
<tr>
<td>σ</td>
<td>7.23</td>
<td>7.15</td>
<td></td>
</tr>
<tr>
<td>Peers Score</td>
<td>54.54</td>
<td>52.98</td>
<td>n.s.</td>
</tr>
<tr>
<td>σ</td>
<td>23.9</td>
<td>24.8</td>
<td></td>
</tr>
</tbody>
</table>
DISCUSSION

Overall, this article suggests a new concept for combining different sources of performance appraisal systems into one score and thus enhancing the ability of managers to assess employees' performance in the organization. The seven-step model can be applied in most organizations that have several relevant sources of performance appraisal available.

Some possible limitations of the present study are the low correlations among the different sources of appraisal. There are several reasons for this phenomenon: a) problems with ambiguous performance appraisal dimensions and items included in the performance appraisal instruments; these shortcomings are relatively simple to resolve by developing clear and uniform performance appraisal instruments; b) problems of different perceptions of the supervisor and the subordinate (supervisors observe different dimensions than peers, peers usually spend more time together, etc.); these problems are not easily overcome.

The authors suggest that future research replicate the model in other organizations. Clear definition of the performance appraisal dimensions will probably lead to better agreement among the sources and will serve the aim of improving the performance appraisal outcomes. Organizations which are suitable for such a study are those in which performance appraisal can be collected from separate sources.

REFERENCES


MANAGEMENT 27 (Fall):273-290.
Harel, G. and Y. Baruch (1992). "The Effect of Educational Background on Performance and Organizational Commitment." Submitted to the HUMAN RESOURCE MANAGEMENT JOURNAL.