Measurement and Assessment Issues in Performance Appraisal

Theresa J.B. Kline  
University of Calgary

Lorne M. Sulsky  
Wilfrid Laurier University

Performance appraisal is a topic that is of both theoretical interest and practical importance. As such, it is one of the most researched topics in industrial and organisational psychology. Several measurement issues are central to performance appraisal including: (a) how performance has been measured, (b) how to improve performance appraisal ratings, (c) what is meant by performance, and (d) how the quality of ratings has been defined. Each of these are discussed along with the shortcomings of the extant literature in helping to come to grips with these important issues. Next, some of the new challenges facing performance appraisal, given its historical focus on single individuals being evaluated, are highlighted. In particular, the appraisal problems inherent in the assessment of team performance and the complexities inherent in multisource feedback systems are covered. We conclude with a short discussion of the litigious issues that can arise as a result of poor performance management practises.

Keywords: performance appraisal, measurement issues

Performance appraisal is a “general heading for a variety of activities through which organisations seek to assess employees and develop their competence, enhance performance and distribute rewards” (Fletcher, 2001, p. 474). Research into this topic has been a major focus of industrial and organisational psychology and management scholars for decades. The dependence of this research area on measurement accuracy is hard to overestimate. Most organisations have formal evaluation systems in place to assess employee performance, but for the majority of those involved, the following quote is apt: “Performance appraisal is a yearly rite of passage in organisations that triggers dread and apprehension in the most experienced, battle hardened manager” (Roberts & Pregitzer, 2007, p. 15). There are many reasons for this dread, and in this article we focus on some of the major measurement issues facing the performance appraisal literature as well as their practical implications.

We start by discussing the historical emphasis on improving the measurement of performance through various format changes and training initiatives. Then we move into what performance appraisal actually means. What exactly do the performance scores represent and more important what should they represent? Then we delve into the psychometric issues of reliability, accuracy, and validity. In the next section of the article we move into some of the relatively newer problems facing performance appraisal researchers and practitioners. Specifically, the assessment of team performance and multisource feedback systems will be presented. We close by highlighting the notion that performance appraisal is not just an ivory tower topic—the outcomes of the process have real meaning and implications for people in their lives. As a result, some performance appraisal problems have made it into the realm of litigation, the results of which we also examine. We close with some thoughts about the future of performance appraisal research and practise.

The Measurement of Work Performance

The history of performance appraisal research is replete with studies examining ways of improving the psychometric quality of performance evaluations rendered by performance raters. Two broad strategies emerged from this literature for maximising rating quality: rating formats and rater training. As we illustrate below, this literature has produced a number of alternative format and training approaches designed to enhance rating effectiveness. However, the meaning of effective ratings has broadened in recent years, while rating format and training research has focused largely on a narrow conceptualisation of effectiveness based on psychometric considerations. We briefly touch on this issue at the conclusion of this section.

Rating Formats

There are at least two important distinctions that can be drawn when examining previous research on alternative rating formats. One concerns the differences between behaviourally based rating formats and graphic, trait-based rating formats (Aguinis, 2009). The other is between absolute and comparative judgements. Each is discussed in turn.

Essentially, behaviourally based scales require the rater to judge either the frequency or the quality of specific employee work actions, whereas trait-based scales require the rater to evaluate the employee on traits (e.g., leadership skills, creativity, etc.) through the observation of employee performance. One example of a behaviourally based scale is the Behavioural Observation Scale (BOS; Latham & Wexley, 1977). The BOS requires appraisers to rate the frequency of specific employee behaviours they observe. An alternative behaviourally based scale is the Behavioural Expectation Scales, usually referred to as a Behaviourally-Anchored Rating Scale (BARS; P. C. Smith & Kendall, 1963). The BARS provides the rater with behavioural expectations associated with alternative scale points (i.e., the “behavioural anchors”). The rater is required to observe employee perfor-
mance and, for a given scale, choose the anchor that best matches or otherwise exemplifies the employee’s observed behaviour.

Graphic-type trait-based rating scales require the appraiser to evaluate the employee on a series of traits or broad competencies. The set of traits/competencies is determined by a job analysis focusing on the underlying skills, abilities, and other characteristics deemed important for performing the job successfully. One immediate challenge for raters is that these scales require the rater to make an inferential leap from behaviour to underlying traits, and this may not be easy to accomplish accurately (Latham, Sulsky, & McDonald, 2008).

Comparative research exploring the relative psychometric merits of these two alternative rating formats has yet to yield any firm conclusions. This is due to at least two reasons. First, there has been a lack of theory guiding a priori predictions about potential scale-based rating differences; therefore, any emerging differences in a given study could be due to measurement artefacts or other methodological issues giving one scale the upper hand in a given study.

Second and, arguably more important, much of this research has relied on indirect indices of rating validity in the form of rater error measures, and these measures of rating quality are problematic from both methodological and conceptual standpoints. (e.g., halo error, leniency error, etc.; see Sulsky & Balzer, 1988). Below, we explore these indices in some detail. For now, suffice it to say that the indirect indicators of validity make it impossible to formulate any definitive conclusions about rating validity (cf. Murphy & Balzer, 1989).

Although we might need to be cautious in drawing any conclusions concerning the relative psychometric merits of alternative rating formats, behaviourally based approaches such as BOS and BARS are widely viewed as superior to trait-based formats from the standpoint of legal defensibility (Latham et al., 2008). This is because the link from behaviourally based scales to on-the-job behaviours is visible and direct. A job analysis can clearly indicate the key job behaviours required for a given job, and these behaviours can be directly represented in either a BOS or BARS rating format. As we noted above, trait-based scales require the appraiser to make inferences from work behaviour to abstract traits such as integrity. This inferential leap is more difficult to challenge from a legal standpoint.

The second important distinction in rating format research is between absolute and comparative judgement rating scales (Aguinis, 2009). Specifically, all of the aforementioned formats require the rater to formulate an absolute performance judgement. However, it is possible to use a format such as a forced distribution format or ranking scale that simply requires the rater to make relative comparisons amongst employees—without the need to assign an absolute rating on a given performance dimension. Wagner and Goffin (1997) argued that it might be easier for raters to accurately evaluate performance in comparative (rather than absolute) terms because social comparisons are a natural by-product of uncertain decision-making situations such as performance appraisal.

One concern arising from comparative rating scales is that it may be difficult for a rater to justify or defend a particular ranking, unless the rater has absolute performance information to bolster the comparative assessments. In addition, although some appraisal decisions require that employees simply be ranked or located on a performance curve (e.g., selecting the top employee for promotion), many appraisal decisions will require an absolute assessment (e.g., calculating bonus pay to the level of performance). Lastly, where appraisals are used for purposes of feedback and development, a comparative assessment will not provide the richness of detail needed to supply the employee with the requisite feedback.

Rater Training

The second broad strategy for improving the psychometric quality of performance ratings is rater training. Three training approaches have dominated previous appraisal research: rater error training (RET), behavioural observation training (BOT), and frame-of-reference training (FOR).

The goal of RET is to introduce raters to common rating errors such as halo error (e.g., rating certain employees based on general impressions), leniency error (producing systematically high ratings across employees), and range restriction error (concentrating ratings on a narrow band of the rating scale; Balzer & Sulsky, 1992; Latham, Wexley, & Pursell, 1975). The assumption in this training is that discussion of these topics will attenuate their effects in the rating process. The approach has been shown to be efficacious for reducing rating errors, (Pursell, Dossett, & Latham, 1980), however, this could have the unintended effect of actually lowering rating validity (Bernardin & Pence, 1980).

For example, if all employees truly are excellent employees and deserve uniformly high evaluations, teaching raters to avoid giving uniformly high ratings could, in fact, inadvertently have an adverse effect on rating quality. In short, RET assumes a normal distribution of performance across employees. Teaching raters to formulate their ratings according to a preordained performance curve will only be efficacious for rating quality if true performance follows the expected performance distribution.

In BOT the goal is to maximise the quality and accuracy of rater observations of employee performance. The roots of this training approach can be traced to cognitive processing models of the appraisal process (e.g., DeNisi, Cafferty, & Meglino, 1984) that delineate how raters process performance information and potentially commit processing errors along the way (e.g., selective or incomplete observations of performance). By training appraisers to avoid processing errors at the point of initial performance observation, this training approach should result in potential improvements in performance rating quality.

Although BOT has not received the same degree of research attention as other approaches, the results are still encouraging. The available studies have suggested that BOT-trained raters produced more accurate ratings compared to others who received minimal or no training (e.g., Hedge & Kavanaugh, 1988; Noonan & Sulsky, 2001; Pulakos, 1986).

FOR is the third major training approach, and is also consistent with cognitive processing models of performance appraisal information (Bernardin, Buckley, Tyler, & Wiese, 2000). The goal of FOR training is to ensure that raters formulate correct impressions concerning employee performance on each performance dimension to be evaluated. Even if raters forget specific details of what a given employee did (or did not do) during the appraisal period, raters should still be able to recall their impressions (cf. Sulsky & Day, 1992, 1994). As long as the impressions are accurate, this should serve to enhance rating quality. FOR training works to calibrate raters so that they agree on (a) the relevance of ratee behaviours for specific performance dimensions, (b) the effectiveness levels of specific behaviours, and (c) the rules for combining individual judgements into a summary evaluation for each performance dimension (Sulsky & Day, 1992, 1994).
In the only published meta-analysis of rater training research, Woehr and Huffcutt (1994) concluded that FOR is the best training approach for increasing rating accuracy with a mean effect size of .83 when compared with other training conditions. A limitation of this research, however, is that most of it was laboratory based (for an exception, see Noonan & Sulsky, 2001). Moreover, to date, there is almost no research critically examining the precise training methods used across these FOR training studies to determine how the training protocol might be either strengthened or streamlined (Chiciro et al., 2004; Sulsky & Kline, 2007).

In concluding this section on rating formats and rater training, it is worth noting that the definition of rating effectiveness at the root of this body of literature is largely based on psychometric issues such as rating validity and accuracy. However, in recent years, there has been a movement to broaden the conceptualization of rating effectiveness, whereby the rating process is assumed to be embedded within a social context (Farr & Jacobs, 2006). Thusly, for example, an “effective” rating might also be one in which the rater or ratee perceives that the rating is fair, or serves to motivate the ratee in intended ways (Levy & Williams, 2004; Murph & Cleveland, 1995; Whiting & Kline, 2007; Whiting, Kline, & Sulsky, 2008).

In sum, we expect there to be a concomitant interest in examining the implications of specific formats and training programs for additional effectiveness criteria such as employee reactions. Although there is some limited work in this area, at least for rating formats (e.g., Tziner & Kopelman, 2002), the possibilities for expanding the evaluative domain of both formats and training is readily apparent in light of this shift toward a more social-contextual framework.

The Meaning of Work Performance

Although the emphasis of previous performance appraisal research has been on issues relating to the measurement of work performance, some performance appraisal researchers also have focused on the issue of conceptualizing the meaning of performance. Without a clear definition of what is meant by performance, the validity of performance ratings derived from measurement scales becomes immediately questionable. Preceding the seminal paper by Austin and Villanova (1992) on the “criterion problem”, a quick foray into the rich history of writing in the area of performance appraisal reveals a number of scholars attempting to grapple with and address the inherent complexities arising when we try to capture the meaning of the ultimate criterion (e.g., Dunnette, 1963; Ronan & Prien, 1971; P. C. Smith, 1976; Wallace, 1965).

Researchers have identified a number of specific and relevant parameters that should underlie any overall theory of performance including: (a) the relevant performance dimensions, (b) the performance expectations associated with alternative performance levels, (c) how (if at all) situational constraints should be weighted when evaluating performance, (d) the number of performance levels or gradients for each performance dimension (see Cardy & Keefe, 1994), and (e) the extent to which performance should be based on absolute versus relative comparison standards (see Wagner & Goffin, 1997).

Although taxonomies of performance have been proposed, they do not consider all the parameters listed above (cf. Bobko, & Colella, 1994). For example, Campbell, M. Cloy, Oppler, and Sager (1993) examined the parameter of relevant performance dimension and presented a taxonomy of dimensions assumed to underlie work in general (e.g., effort and personal discipline). Further work has been conducted to examine and refine the Campbell et al. formulation (e.g., Tubre, Arthur, & Bennett, 2006). Although we will not review the Campbell et al. theory here, it is a step in the right direction even though it does not consider all the possible parameters. Most important, the intent of the theory was to develop a conceptualization of performance that generalizes across jobs and situations. There are other notable taxonomies that have provided a useful basis for defining the content of work performance (see Tubre et al., 2006).

Arguably, much of the theory and research concerning the development of conceptualizations of performance have focused on two parameters: The delineation of performance dimensions and expected associated performance standards. Two methods to identify the parameters include a bottom-up process based on the critical-incidents technique, and a top-down process based on competency modeling. The first approach advocated for identifying performance dimensions and standards is inductively based and grounded in the use of the critical incidents technique (Flanagan, 1954). The critical incidents method was not developed specifically for developing performance theories, but it does provide an inductive, bottom-up process to the identification of performance dimensions and performance standards development.

Using this approach, subject-matter experts (SMEs) write exemplars of performance at various effectiveness levels. These exemplars or incidents are sorted into dimensional categories, and a given incident “survives” to later stages of the process if the vast majority of SMEs classify it into the same dimension. If a BES or BARS is to be developed, the incidents are rated for effectiveness level, and the mean effectiveness rating across SMEs serves as the de facto performance level for the incident. Some of these incidents can then be transformed into behavioural anchors (for BARS; those incidents with a small standard deviation for the effectiveness ratings) or can be compiled to give both raters and ratees a list of performance expectations.

Sulsky and Kown (1998) observed that this approach is limited in a number of ways. First, there is no a priori or “top-down” theory to guide development, and the process relies on statistical considerations including averaging—which can mask apparent disagreements across SMEs. Further, the procedure does not shed light on how or whether dimensions should be optimally weighted to arrive at an overall performance assessment. Finally, important dimensions of the criterion space may be omitted if incidents are not written for these dimensions.

A second deductively based approach that may be used for identification of performance dimensions and associated performance standards invokes a top-down approach through (a) the identification of core competencies that generalise across jobs, and (b) the identification of behaviours and associated standards for specific jobs in light of the competencies (e.g., Fletcher & Perry, 2001; Smither, 1998). In sum, this involves developing a performance appraisal system defined by behaviourally based core competencies that are common across jobs within an organisation.

A competency model is developed from a review of functional job analysis data and a content validation process (Smither, 1998). First, core competencies are identified that cover the majority of positions and reflect the organisation’s strategic goals. The competencies are then defined at the behavioural level and including criteria for differentiating between different levels of expertise (Fletcher & Perry, 2001; Smither, 1998).
Although the competencies remain the same across all positions, the behavioural expectations of individuals who fill those positions vary with their level of responsibility. Smither argued that a common competency model (based on an organisation’s strategy) can guide and integrate numerous human resource processes.

One potential criticism of both the inductive and deductive approaches described above is that they capture only some aspects of the job. In particular, several researchers have suggested that job performance relates to two distinct sets of behaviour—those that are defined in the formal job description and those that are defined by the organisation’s social context (Murphy & Cleveland, 1995). For example, Borman and Motowidlo (1993) discriminated between task (proficiency at core technical activities) and contextual (behaviour that contributes to the organisational, social, and psychological environment in accomplishing goals). Contextual behaviours include such behaviours as volunteering, helping, and endorsing organisational objectives, and have been shown to be an important aspect of effective performance.

Another construct reflecting discretionary work behaviour, organisational citizenship behaviour, (OCB) has received far more research attention compared to contextual performance (B. J. Hoffman, Blair, Meriac & Woehr, 2007). Hoffman et al. noted that recent conceptualizations of OCB are identical in meaning to contextual performance, and that OCB relates more strongly to work attitudes (e.g., job satisfaction) compared to task performance. Although there have been attempts to uncover the dimensiolity of OCB (e.g., a two dimensional model whereby OCB is directed at individuals or at organisations as a whole—see Williams & Anderson, 1991), their analysis of the OCB construct suggests that OCB is a unitary latent construct.

A second criticism of both the inductive and deductive approaches for identifying performance dimensions and associated standards is that organisations are increasingly characterised by complexity and continual change, which affects the nature of work itself and the meaning of good performance (Fletcher & Perry, 2001). In an ever-changing work environment, the definition of jobs and what is meant by good performance is less stable, and hence more elusive. In addition, important aspects of job performance are context specific—meaning that the same job may not have identical duties and responsibilities.

Given the multifaceted nature of work and all of the associated complexities just described, the question remains whether it is possible to adequately define performance dimensions and their associated standards. Of course, the answer to this question will depend on the meaning of adequate. Despite all the complexities inherent in conceptualizing the relevant performance dimensions and performance expectations relating to those dimensions, a combination of both bottom-up and top-down approaches may be the best strategy for capturing the meaning of performance.

Although this hybrid approach may be promising, some of the intricacies such as how, if at all, to incorporate contextual performance, will still need to be considered. In addition, to really hone in on what is meant by performance, we will need additional strategies and approaches to address additional parameters (e.g., how to factor in situational constraints on performance) not specifically addressed by either the bottom-up or top-down process.

Examining Rating Quality

As we noted earlier, the hallmark of previous rating format and rater training research has been to discover ways of enhancing the psychometric quality of performance ratings (Murphy & Cleveland, 1995). To this end, research has emphasised indirect indices of rating validity as surrogate measures to more direct validity indices. The two primary indirect approaches adopted in previous performance appraisal research are (a) rater error measures, and (b) measures of rating accuracy. Below we examine these two categories and discuss some of the limitations arising from their adoption.

First, rater error measures became popular in the 1970s as indirect indices of rating quality (Saal, Downey, & Lahey, 1980). The idea was that ratings devoid of such errors as halo error, leniency/severity error, and central tendency/range restriction error would necessarily be higher in validity compared to ratings suffering from one or more of these errors. However, it became clear over the years that these errors may not signal that the ratings are suffering in psychometric integrity. As we described earlier in connection with the RET approach to rater training, discovering these errors may or may not suggest a rating validity problem. If, for instance, a ratee deserves high ratings across performance dimensions, the ostensible existence of halo error (sometimes operationalized as a low standard deviation in ratings for a given ratee, see Balzer & Sulsky, 1992) in fact signals that the ratings possess validity for that ratee. Not surprisingly, it has been shown that these errors do not predict rating accuracy (Murphy & Balzer, 1989; Sulsky & Balzer, 1988). Because of the inherent limitation of these traditional error measures as indirect indices of rating validity, Murphy and Cleveland (1995), amongst others, recommended that these indices be abandoned as surrogate indices for more direct indices of rating quality.

The second common approach for assessing rating validity is indirectly through the assessment of rating accuracy. Here, the idea is that accurate ratings must necessarily be high in validity because accuracy presumes validity—much the same way validity presumes reliability (Kline, 2005; Sulsky & Balzer, 1988). Thusly, for example, it makes little sense to ask if a thermometer is properly calibrated and giving accurate temperature readings (e.g., is the reading off by a constant amount of 10 degrees?)—unless it is already determined that the thermometer is valid for the purpose of determining the temperature in the first place.

There are a number of operational definitions of rating accuracy, although the Cronbach (1955) component accuracy scores have been the most popular accuracy indices in previous appraisal research. Essentially, the Cronbach accuracy scores examine the numerical distance between a set of ratings produced by a given rater, and another set of ratings provided by expert raters (variously termed “true”, “target”, or “comparison” scores). The closer the rater’s ratings are to the true score ratings the more accurate the ratings are assumed to be (Sulsky & Balzer, 1988).

What makes the Cronbach indices useful, however, is that they decompose the rating distance into four orthogonal components: (a) elevation accuracy, (b) differential elevation accuracy, (c) stereotype accuracy, and (d) differential accuracy. Elevation accuracy examines whether the overall level of a rater’s ratings is different from the overall level of the target ratings. Thusly, this index is useful for diagnosing the possibility that the rater is being overly lenient or harsh relative to the intended expert ratings. For a complete descrip-
The important point here, however, is that this finer grained decomposition of the overall ratings distance allows the researcher to choose the index (or indices) most relevant to the research questions at hand. Thusly, for example, if it is hypothesised that a given manipulation should cause raters to overly inflate their ratings, elevation accuracy becomes an appropriate index to examine as the criterion of rating quality.1

Although it can be argued that accuracy measures represent a decided improvement over the traditional rater error measures, accuracy measures are not without their own problems. For instance, Sulsky and Balzer (1988) noted that the quality of the true scores must be established if it is to be argued that accurate ratings are deemed to be ratings numerically close to the true ratings. Moreover, the quality of the true ratings is only going to be as good as the technologies used to identify them—and there are potential problems with the approaches taken to develop these ratings in some of the previous appraisal literature (Sulsky & Balzer, 1988).

Perhaps an even more challenging issue for accuracy measures is that they can be difficult, if not impossible, to obtain in field settings, unless there is some way for expert raters to observe and capture the full array of rater performance (Sulsky & Balzer, 1988). In some contexts (e.g., assessment centres) this task is easier compared to others (e.g., performance appraisal in which rater performance must be sampled over a considerable time period).

Yet another issue that challenges the potential utility of accuracy measures is the distributional properties of the true scores. For instance, assume there are two sets of these true scores for a seven-point rating scale (with one rater and three performance dimensions): Set A contains the ratings 4, 5, 4, whereas Set B contains the ratings 1, 3, 2. Also assume a study is designed whereby it is assumed that raters will inflate ratings under certain circumstances (e.g., the ratings will be used for reward purposes). If elevation accuracy is chosen as the accuracy criterion, it becomes evident that Set B allows for a greater range of elevation scores compared to Set A. Therefore, statistical power will be enhanced if Set B is used, and relatively attenuated if Set A is adopted. It can also be shown that manipulating the mean and variance of these scores can affect which of the accuracy component scores is most likely to be affected by a given rating manipulation. We are not aware of any studies whereby the distributional properties of these true scores are explored as part of the process of developing the accuracy measures. In sum, the failure to obtain an expected effect in a given study could be partly or wholly explainable based on the properties of the true scores used in the computation of rating accuracy.

Finally, it is important to remind ourselves that validity is not a property of a set of ratings—it is the inferences we draw from the ratings that will either be valid or not (Society for Industrial and Organizational Psychology, 2003). Consider a situation in which there are two employees, and they each truly deserve the following ratings for three performance dimensions: Ratee A: 7, 6, 7 and Ratee B: 6, 5, 5. Assume Rater A produces ratings of 5, 5, 6, and 6, 7, 7 for Ratees A and B, respectively, and Rater B produces ratings of 3, 2, 2, and 2, 1, 1, for the two respective ratees. Clearly, Rater A is more accurate according to our conceptualization of rating accuracy. However, what if the appraisal decision is to select and promote the top performing ratee? Rater B would formulate the correct inference that Ratee B is superior, whereas Rater A would not. Therefore, paradoxically, Rater A is seemingly more accurate, yet it is Rater B who formulated a correct inference concerning who is the best person to promote.

The lesson from this contrived example is that we must always be mindful of the inference(s) we wish to draw from our performance ratings when interpreting any index of rating quality—including rating accuracy. To be fair, the likelihood is that in most situations, rating accuracy will be positively correlated with decision inference quality. Nonetheless, the example serves to emphasise that accuracy is an indirect indicator of rating validity (when validity is defined in terms of validity of inferences), and we should not just assume that accuracy necessarily translates in validity.

**Performance Appraisal and the Social Context**

Perhaps even more important, the above example reminds us that if validity is an inference, we might decide that our desired inference has little or nothing to do with psychometric issues. Suppose that the goal of a rater is to produce a set of ratings perceived as fair across employees given that the thrust of appraisal research has increasingly focused on the social context of performance evaluations (e.g., Levy & Williams, 2004). Thusly, perhaps from the standpoint of the rater, a "valid" set of ratings are those that allow us to correctly infer that employees agree and accept that the process undertaken by the rater to generate the ratings was fair and unbiased.

By viewing performance appraisal from a social-contextual perspective, it may cause us to reconsider, or at least broaden, the meaning of rating effectiveness. In fact, from the rater’s perspective, the goal may be to motivate employees to improve, and producing artificially low ratings may be a motivated attempt on the part of the rater to accomplish this objective (M urphy & Cleveland, 1995). If the ratings indeed serve as a motivational booster, can we say the ratings are "valid"? Ultimately, it may depend on the inference(s) to be made, or goals intended, based on a set of performance ratings. Levy and Williams (2004) suggested many social-contextual variables that may play a role in determining the utility of a given system in a given context including, but not limited to: organisational culture, legal climate, trust, rater training, and appraisal documentation. Their long list of variables highlighted the complexity of the performance appraisal system that goes far beyond the psychometric properties of the rating instruments.

This concludes the portion of our discussion about the theoretical and statistical issues inherent in performance appraisal measures. We turn now to some of the areas in which innovation in performance appraisal research is needed as it is applied in organisations. As these sections are described, the importance of ensuring high quality evaluation mechanisms and rater training is obvious. In addition, though, the relevance of the social context of the performance appraisal system will be shown to be at least as pertinent.

**Team Performance Appraisal**

There are two issues that continue to plague the team performance appraisal literature: “How do we measure team performance?” and...
“How do we use the measures of team performance?” There are some subtle but important issues in team performance appraisal that warrant particular attention by researchers and practitioners. Each question will be examined as to the consensus in the literature and the implications for researchers and practitioners suggested.

The first question: “How do we measure team performance?” remains elusive. As noted earlier in this paper, much of the past literature on performance appraisal has focused on how to develop and validate appropriate measures. Although not discounting this substantial work, developing an appraisal metric for teams poses problems not encountered in the literature on single employees. Part of the reason for this is that teams are unique; some are formed specifically to deal with a single project, some are long-standing teams that perform a function in an organisation (e.g., technology support or clerical support), and some teams come together for repeated performances (e.g., fire fighters, surgical, flight crews; Kline, 1999). The search to find a standardised set of items that assesses team performance across organisations is rather futile. Instead, it is suggested that a standardised process to design an assessment system for each team makes more sense (Joness & Schilling, 2000). This process includes identifying the role of the team in the organisation and analysing the tasks of the team. A more detailed explanation follows.

First, the purpose of the team needs to be established. This is found by asking relevant stakeholders their views of the team’s outcomes. Managers and customers are excellent sources to provide input to this process. Managers can assist in making sure that the business strategy of the organisation is carefully conveyed to the team so that the team’s goals can be aligned with the organisation’s goals. They can also provide useful information about what the team is going to be assessed on from the organisation’s perspective. Clients or customers provide information about the expected quality and quantity of the team’s outcomes. Team members are also a good source of information about the team’s work. Members with experience can be helpful in describing what has worked well in the past and what has not in terms of the team’s performance. Fresh ideas about what the team should be focusing on can come from other newer team members. Other individuals or teams within the organisation provide yet another perspective on a team’s outputs. These are especially useful if the team’s outputs are inputs into another work process.

Given this extensive list of contributors to team performance, it suggests several points at which researchers and practitioners can be helpful to teams. Creating easy-to-use inventories for the stakeholders to complete is a task that researchers are particularly well-equipped to carry out. They are also good at determining which of the items in those inventories are most reliable and useful. Understanding why the various perspectives are unique or similar and the underlying mechanisms involved in forming these perspectives would be the basis for a substantial research program. Practitioners working with teams need to have this information to effectively facilitate the team’s work. They are in a position to gather this information, collate it, synthesise it, and report back to the team on the findings. They can be quite helpful to teams in using this information to set goals and monitor performance.

We have just presented an approach that would likely provide very idiosyncratic assessments of team performance. In fact, many researchers have conducted a process similar to that describes. They have found that despite the fact that teams are unique, there seem to be some common themes on which most teams should be evaluated (e.g., Brannick, Salas, & Prince, 1997; MacBryde & Mendibil, 2003). These are generally divided into two primary categories: team processes and outcomes. A brief description of these findings will provide researchers and practitioners a starting point in their respective work on team assessment.

Team processes include matters such as how well: (a) the team makes decisions, (b) the team communicates with each other, (c) the team gives and receives feedback, (d) the team demonstrates leadership, and (e) the team members’ attitudes toward each other and their task. This list is by no means exhaustive, but identifies potential sources of input for team performance process assessment. Team outcomes usually centre on the quality and quantity of goods or services provided by the team. Often the capability of the team to complete their work on time and within specified budgets is important to assess. Finally, team attitudes as an outcome have been cited as quite relevant (Sundstrom, DeMeuse & Futrell, 1990). Team members who do not want to work together in the future pose problems for organisations and so member attitudes about the team are important in and of themselves.

The next question is: “How do we use the measures of team performance?” At the individual level, performance is often used to make decisions about salary increments, promotions or terminations, and training needs. This approach does not translate easily to uses for teams. For example, it is not typical that teams as a unit are promoted or fired. However, performance measures can and should be used to identify areas of strength and weakness in a team’s performance (Brannick et al., 1997). In this regard the use is similar to that of individual-level performance appraisal. One particularly problematic area in terms of human resources practises is the degree to which an employee’s salary is tied to their team’s performance.

It has been known for quite some time that tying the team members’ compensation to the team’s performance enables optimal performance (e.g., Geber, 1995). However, only one fourth to one third of firms using teams actually engage in creating a team-based performance management system (Pastrenak, 1994). J. R. Hoffman and Rogelberg (1998) provided examples of seven different team incentive systems including, (a) profit sharing, (b) goal-based incentive systems, (c) discretionary bonus systems, (d) skill incentive systems, (e) member skill incentive systems, (f) member goal incentive systems, and (g) member merit incentive systems. Note that the last three actually focus on individual team members rather than the team as a whole.

There is little research in this area and few guidelines for practitioners. However, there is clearly a need to determine the optimal ratio of team-based to individual-based compensation. We might speculate as to the ratio that would be expected, however, by examining some cross-sector data. For example, the average variable pay rates in Canada in 2007 were 6.4% of base pay (Conference Board of Canada, 2008). Thusly, variable incentive systems based on team performance would likely need to be at least this magnitude to be salient to team members.

A next step would be to identify variables that predict team performance at different ratios of team to individual compensation (e.g., 1:9, 2:8, 3:7, etc.). This will assist in understanding team performance and how it is related to pay systems. For example, there have been variables suggested that should lead to different ratios such as the level of team interdependence, the degree of shared goals, and various member contributions (e.g., Zingheim & Schuster, 2007). However, what is not known how these variables predict team performance.
Clearly, there is work to be done on both the practitioner side in terms of facilitating team performance with incentive systems and on the researcher side in terms of building a model of team performance that includes team incentive systems as a primary variable of interest.

Consistent with our earlier comments about the social context and its importance in interpreting and using team performance measures, the literature suggests that the feedback culture of the organisation is extremely important in making proper use of performance appraisals for teams (London, 2003). A positive feedback culture is one in which all parties feel comfortable in both providing and receiving feedback. Although arguably this is important for single-employee appraisals, teams are appraised not only by their managers, but often by each other as well. Developing a positive feedback culture in which team members are trained to provide effective feedback is key to developing high-performance teams.

Performance appraisal measurement and assessment and the degree to which it is perceived to be a valid and reliable process clearly have implications for team performance. Although the issues discussed earlier in this article do clearly have an impact on performance management in the individual level, performance assessment, and management at the team level provides a unique set of challenges for both researchers and practitioners.

**Multisource Issues in Performance Appraisal**

Performance appraisal systems that are set up to gather information from a variety of sources, in addition to traditional supervisory input, are called multisource (commonly labelled 360-degree feedback) programs. These programs are common, with 43% of organisations reporting that they use at least some form of multisource system (Brutus & Derayeh, 2002). It has been known for quite some time that performance feedback from multiple sources, including self, subordinates, customers, and peers, has been shown to lead to more reliable ratings, better performance information, and greater performance improvements than traditional performance appraisal methods (Dominick, Reilly, & McGourty, 1997; Facteau, Facteau, Schole, Russell, & Poteet, 1998; Latham & Wexley, 1993). The relationship between the implementation of a multisource system and performance improvement, however, is not just a simple, positive, bivariate phenomenon.

For example, Flint (1999) found that if employees' performance ratings were low in a multisource system, their performance improved only if they perceived the process to be a fair one. If it was perceived to be unfair then performance actually decreased. Smith, London, and Richmond (2005) found in a longitudinal study of military leaders, that individuals high on emotional stability were rated as more likely to use feedback from multiple sources and leaders high on responsibility were more likely to be rated as more motivated to use the feedback.

One of the most important moderating factors regarding the utility of providing feedback to individuals is simply whether the feedback is accepted (Ilgen, Fisher, & Taylor, 1979). Thusly, although a multisource approach to performance assessment makes sense in theory insofar as multiple measures usually provide for a more reliable and valid assessment of a phenomenon (Campbell & Fiske, 1959), in practice, the source of the data, what is being measured, how it is fed back to the person being evaluated, and characteristics of the person being evaluated all come into play to make this a less than straight-forward process. A gain, the importance of the social context of the system plays a substantial role in its effectiveness. Equivocal findings regarding the use of multisource systems confirm that something other than the accuracy of the ratings is operating to influence performance (e.g., Seffert, Y ukl, & M cDonal d, 2003). For example, the overall perceived fairness of the system by the ratees is critical to its success (Facteau & Facteau, 1998).

In the next sections, we will examine the literature on the most prevalent sources of performance assessment and how this has an impact of research and practise. In particular the sources are, (a) self-ratings, (b) peer ratings, and (c) subordinate ratings. The intercorrelations between these ratings are notoriously low (M abe & West, 1982). Thusly, it is important to understand what may influence the low agreement levels as well as why they might legitimately occur and not be a source of measurement error.

Self-ratings of performance are perhaps the most common of the "other" sources of performance feedback that have been in use for some 50 years (B assett & M eyer, 1968). The primary problem with self-ratings is that they frequently disagree with supervisory ratings and they differ in the expected direction—the employee rates themselves higher than does the supervisor (e.g., T s u & B arry, 1986). These discrepancies can foster a useful dialogue between supervisor and employee. They can also highlight that the employee may not be aware of the goals of the organisation. Campbell and Lee's (1988) model posits reasons for the discrepancies including the following: (a) informational differences about what is to be performed and how, (b) different schemas associated with employee performance, and (c) psychological defences by the employee about their performance. Bono and Colbert (2005) found that congruent self and other evaluations led to high levels of satisfaction, but not necessarily to better performance. The moderating roles of goal commitment and core self-evaluations are cited as playing a relevant role in the use of performance assessments. Self-ratings then, are of limited use as a general strategy given the number of moderating variables in the relationship between self-ratings and improved performance.

Peer ratings are again, unfortunately highly unreliable (Love, 1991). This may be due to peers having access to different types of information about the ratee. For example, one peer may interact with an employee as a teaching colleague and another with that same employee as a research colleague. When they are asked about the performance of the ratee, the peers are likely to focus on quite different information when making their evaluations. Despite this Shore, Shore, and Thornton (1992) found that peer ratings were superior in predicting performance than were self evaluations. In addition, Harris and Schaubrooke (1988) found the relationship between supervisor and peer ratings to be higher than self with supervisor or self with peer ratings. However, much of the evidence suggests that peer ratings have low user acceptance levels (e.g., C ederblom & L ownsbury, 1980). A study by F arh, C annella, and B edean (1991) found that users were more likely to accept peer feedback if it was for developmental purposes only, and when they were being used for developmental purposes the reliability and validity of the ratings was higher than when used for evaluative purposes. This study highlights the complexity of the multisource approach to performance assessment. That is, although the ratings themselves may be susceptible to psychometric problems, a much more relevant issue is the perceived quality and usefulness of the ratings by employees. This issue is not amenable to simply improving the measure of...
The amount of agreement about performance requires some research. To get the system right, both from measurement and implementation, delivering the feedback from all the sources requires ineffective system (Kaplan, 1993). Individuals are being evaluated on particularly if a rater provides negative feedback. Proper following work. Confidentiality of the source may also become relevant, would be spending more time evaluating work than actually conducting. How to combine the information onto a coherent package that generates. Peers would obviously interact with an employee in a much different manner than supervisors or subordinates. Identifying what performance dimensions are to be rated by which source is also complex. Peers would obviously interact with an employee in a much different manner than supervisors or subordinates. How to combine the information onto a coherent package that an employee could actually use to improve their performance is also a nontrivial task.

The issues just cited are the administrative drawbacks to such a system. A nother clear problem is that if everyone is evaluating themselves, their peers, their subordinates, and their supervisors as well as collecting performance information from their customers, it is easy to see how the system could take on a life of its own and employees would be spending more time evaluating work than actually conducting work. Confidentiality of the source may also become relevant, particularly if a rater provides negative feedback. Proper follow-through to assess if changes have occurred is a hallmark of an effective system (Kaplan, 1993). Individuals are being evaluated on their performance about which people are usually quite sensitive. Therefore, delivering the feedback from all the sources requires interpersonal skill. All of these factors combine to suggest that introducing multisource feedback systems should be done judiciously and with the assumption that much time and energy will need to be spent to get the system right, both from measurement and implementation perspectives.

There are several aspects of multisource performance systems that still require research to further our understanding of the phenomenon. The amount of agreement about performance requires some research. Although it is expected that high levels of agreement enhances individual and organizational outcomes, and low agreement results in poor outcomes (Yammarino & A brwater, 1993), under what circumstances it is appropriate to generate more agreement between parties and how to do so is not clear. There also seems to be a movement in the literature to focus more on the individual differences of the ratee in terms of acting on the multisource performance feedback. Although some of these have been identified, this research area is still quite new and provides several opportunities for future studies. Other potential moderators would include the organizational context, degree of trust between raters and ratees, and the use to which the performance evaluations are put. This is a quite complicated area of performance appraisal given the large number of potential variables involved in trying to understand and best utilize such a system.

Litigation Issues in Performance Appraisal

Performance evaluations are used to make personnel decisions in addition to providing developmental feedback to the employee. If an employee feels that a decision has been made based on the evaluation, there are several options that can be pursued. The most commonly used approach is to appeal the decision if such a system exists. This approach is the least adversarial and allows the ratee to put forward arguments as to why the performance appraisal is not an accurate assessment of performance. The next approach to use should the appeal not be upheld is through a formal grievance procedure. These procedures are usually part of any unit that has collective bargaining privileges. It is of low cost because it usually involves the employee’s time, administrative time in the organization, and perhaps union official time. The next level is arbitration. This is a process used to resolve disputes by the intervention of a third party agreed on by the disputing parties or provided by a legal body. This is a much more expensive option insofar as fees for lawyers, an arbitrator, expert witnesses, and all of the accompanying expenses that are involved. Finally, an employee may bring forward a lawsuit against an employer for wrongful dismissal based at least in part by a flawed performance evaluation process. This is by far the most expensive of the options as court time and much more lawyer and expert witness time will accrue over the course of the lawsuit. The social-contextual variable of the legal climate of both the organization and wider society is suggested to play a role in whether or not the ratee engages in grievance procedures over a performance appraisal decision (Levy & Williams, 2004).

S. Smith (2008), an attorney in the United States, provided an excellent overview of the importance of using performance appraisals properly—especially as an antidote to litigious activity on the part of disgruntled employees. She noted that although most large companies have a formal review process, many smaller companies leave it to the individual managers to decide where, when, and how to conduct performance evaluations. Employers who terminate employees without documenting and communicating their performance deficiencies run the risk of negative legal decisions and may have to pay out large monetary awards to these employees. Smith also argued that there should be consistency measurement for individuals within the same job family and performance standards should be clearly understood by all parties. There should be clear statements as to what constitutes excellent, good, fair, poor performance on the relevant dimension.
Bacal and Associates (2008), a Canadian law firm, stresses that the numerical ratings or rankings provided by many performance management systems have the allure of seeming to be objective because there are numbers attached to the ratings. They pointed out that the numbers, though, were based on subjective perceptions and should thusly be heavily scrutinized for their consistency and validity. In an Alberta Labour Relations Board ruling (Asbell, 2004), the Board agreed with the union’s contention that the performance appraisal was not an accurate reflection of the plaintiff’s work. They recommended that the employer remove the performance appraisal for the disputed period from the plaintiff’s personnel file. Arbitrator Bowman (Armstrong, Carpenter, Kline, & Magennis, 1996) in a case in Manitoba stated,

There is no question that this article introduces what is commonly referred to as a “threshold clause” for promotion or hiring. This means that where there is more than one candidate for the position the candidates are not compared with each other, for the purpose of determining which may be the most skilled, or most qualified, but rather, that each is compared to an objective standard indicating the requirements for successfully carrying out the position. Of the persons who demonstrate capacity to meet the position’s requirements, with perhaps some limited training and generally some familiarization time, the most senior will be given the position over other qualified applicants. In other words, amongst those who can do the job, the most senior is entitled to receive it. (p. 136)

Another example comes from the personal experience of one of the authors. In this instance the performance evaluation was lower than anticipated. The unit had recently undergone a revision of its performance evaluation system whereby different dimensions of performance were going to be substantially altered from that experienced for the past many years. There was a decision to use the new weighting system for the current round of performance evaluations (to take place less than 2 months after the decision to change the system had been taken). The author formally appealed the decision, basing the argument on the fact that the new system had not been in place long enough to have had an impact on the performance dimensions and that the long-standing system that had been in place should have been used to assess performance.

All of these examples suggest that attention needs to be paid to how performance is evaluated and how those performance measures are to be used. Because of the potential liability inherent in performance assessment, it is probably a good idea to conceptualize the performance appraisal system and the overall evaluation that ensues from that system as a general “validation” issue. If one does, it opens the door to using such documents as the system—not just the measurements but how the evaluation is conducted and what the information is used for—are important considerations for organizations. Careful design and ongoing evaluation of the system itself can go a long way toward making sure that performance assessments are perceived as fair and justified.

Summary

We have tried to provide a balanced perspective in this article by capturing some of the more fine-grained aspects of performance measurement issues (e.g., rater training proficiency, rating accuracy) as well as some of the more macrolevel issues (e.g., meaning of performance, validity of inferences). In addition, we have provided a couple of examples of the true relevance of this topic to organizations today (team performance assessment, multisource assessment, and litigation). Given its importance to the everyday life of working people, it is safe to assume that performance appraisal is, and will continue to be, a prominent feature in the landscape of industrial/organisational psychology research and practice.

Résumé

L’évaluation de la performance est un sujet intéressant sur le plan théorique et important dans la pratique. De fait, il s’agit d’un des sujets les plus étudiés en psychologie industrielle et organisationnelle. Plusieurs questions liées à la mesure sont centrales en évaluation de la performance, dont : (a) comment la performance a été mesurée, (b) comment améliorer la cotation dans l’évaluation de la performance, (c) que signifie performance et (d) comment la qualité de la cotation a été déterminée. Chacune de celles-ci est abordée à la lumière des limites de la littérature existante, de manière à en approfondir la compréhension. Ensuite, quelques-uns des défis que pose l’évaluation de la performance, compte tenu de la tendance historique à mettre l’accent sur l’évaluation individuelle, sont soulignés. Particulièrement, les problèmes d’évaluation inhérents à l’évaluation de la performance d’équipe et les complexités inhérentes aux systèmes de rétroaction multisources sont couverts. Nous concluons avec une brève discussion à propos des questions litigieuses pouvant découler de pratiques managériales inadéquates.

Mots-clés : évaluation de la performance, questions liées à la mesure

References


