16

Job Performance: Assessment Issues in Personnel Selection

Chockalingam Viswesvaran and Deniz S. Ones

An important construct in Industrial, Work and Organizational (IWO) psychology, organizational behavior, and human resources management (personnel selection, training, and performance evaluation) in general, and personnel selection in particular, is the construct of job performance. Job performance is the most important dependent variable in IWO psychology (Schmidt & Hunter, 1992). A general definition of the construct of job performance reflects behaviors (both visually observable and non-observable) that can be evaluated (Viswesvaran, Ones, & Schmidt, 1996). In other words, job performance refers to scalable actions, behaviors, and outcomes that employees engage in or bring about that are linked with and contribute to organizational goals (Viswesvaran & Ones, 2000). To date, most researchers focusing on the construct of job performance have confined themselves to particular situations and settings with no attempt to generalize their findings. Also, there has been an emphasis on prediction and practical application rather than explanation and theory building. The consequence of these two trends has been a proliferation of the various measures of job performance in the extant literature. Virtually every measurable individual differences dimension thought to be relevant to the productivity, efficiency, or profitability of the unit or organization has been used as a measure of job performance. Absenteeism, productivity ratings, violence on the job, and teamwork ratings are some examples of the variety of measures used to measure job performance.

There are multiple uses for job performance data. In selection contexts, measures of job performance are used to validate predictors. Thus, the choice of the job performance measure has important substantive implications for our practice and science of personnel selection. Measures of individual job performance play a central role at each step of the personnel selection function. Consider the first step in selection: recruitment of qualified applicants. One question in recruitment is whether the different sources of recruitment result in attraction of individuals who differ in job performance levels (Barber, 1998). Following successful recruitment efforts, attempts are made to identify individual differences variables that are related to individual differences in job performance, and select individuals based on those characteristics (Guion, 1998). Individual differences in job performance are assessed and those assessments are used in placement and promotion decisions.

The Blackwell Handbook of Personnel Selection, First Edition.
Edited by Arne Evers, Neil Anderson and Olga Voskuijl.

Individual job performance data can be used in numerous other ways that have relevance for personnel staffing. Cleveland, Murphy, and Williams (1989) identified several uses of individual job performance data. They classified these uses into four categories: 1) between-person decisions, 2) within-person decisions, 3) systems maintenance, and 4) documentation. The need for clear documentation of individual job performance is evident in several landmark legal decisions (Malos, 1998). Individual job performance assessment has been used for administrative purposes for many decades (Whisler & Harper, 1962). DeVries, Morrison, Shullman, and Gerlach (1986) report that surveys conducted in the 1970s in both the USA and the UK indicated the prevalence of individual job performance assessment for the purpose of making administrative decisions. Thus, understanding the issues in job performance assessment is critical in personnel selection.

Several issues have been raised in the assessment of job performance (Austin & Villanova, 1992; Campbell, 1990; Viswesvaran, Schmidt, & Ones, 2002). Although the literature on each of these issues can be expanded into a book-length exposition, in this chapter we will attempt to cover some of the major issues. First, we will briefly discuss the different measurement methods by which job performance could be measured. Second, we will provide an overview of different sources of ratings, the widespread use of multisource feedback, and the claims made for the unique perspective of raters at different levels of an organization. Our review found scant empirical evidence that raters at different levels of an organization have different conceptualization of job performance dimensions. When ratings are provided on a job performance dimension, say interpersonal competence or teamwork, supervisors, peers, and subordinates of an employee provide equivalent construct-level ratings. Different manifestations of the construct could be the focus for different sources, but the underlying construct assessed remains the same.

Following this, we will discuss the content or dimensionality of job performance measures. The relative merits of using a broad or a narrow conceptualization for assessing job performance and the subsequent implications for issues such as differential validity of predictors are covered. Finally, we conclude this chapter with some emerging issues, both psychometric (e.g., appropriate reliability coefficient to use, role of halo) and substantive (e.g., assessing team performance).

Measurement Methods

Individual job performance can be measured utilizing different methods. However, these methods can be classified into two broad categories: 1) organizational records, and 2) subjective evaluations. Organizational records are considered to be more "objective," in contrast to the subjective evaluations that depend on a human judgment. This distinction is parallel to what Smith (1976) described as hard criteria (i.e., organizational records) and soft criteria (i.e., subjective evaluations).

However, it is important to stress that even "objective" organizational records depend on human evaluation/judgment and recording of observed events. Furthermore, it is not necessarily the subjectivity that should be cause for concern. Measurements of job performance should be evaluated on psychometric properties such as criterion contamination, deficiency, relevance, reliability, appropriateness, etc. Blum and Naylor (1968) identify

eleven dimensions or characteristics on which the different criteria can be evaluated, whereas Brogden (1946) identifies relevance, reliability, and practicality as the criteria for criteria. Relevance refers to the overlap between what is measured and what was intended to be measured. Thus, criterion relevance is similar to construct validity of measures. Criterion contamination is the inclusion of sources of variance in a measure that was not intended in the theoretical conceptualization. Criterion deficiency is the lack of relevant sources of variance in the measure employed but which was intended to be measured (i.e., the intended criterion). Practicality refers to the ease with which a measure could be administered. On all these criteria (e.g., contamination, deficiency, etc.) both organizational records and rater judgments and evaluations are affected to different degrees. As such, there is no basis for assuming that organizational records are more "objective" or "accurate" than ratings.

Methods of assessments should be distinguished from types of criteria. Different types of criteria have been formed based on either the time span of performance considered or what is included in the measure of performance. For example, Thorndike (1949) identifies three types of criteria: immediate, intermediate, and ultimate criteria. The ultimate criterion summarizes the total worth of the individual to the organization over the entire career span. The immediate criterion is a measure of individual job performance at that particular point in time. Intermediate criteria summarize performance over a period of time. Similarly, Mace (1935) argued that measures of individual job performance can stress either capacity or will to perform. This distinction is a forerunner to the distinction between maximal and typical performance measures (e.g., DuBois, Sackett, Zedeck, & Fogli, 1993; Klehe & Anderson, Chapter 15, this volume; Sackett, Zedeck, & Fogli, 1988). Maximal performance is what an individual can do if highly motivated, whereas typical performance is what an individual is likely to do in a typical day. The distinction between ultimate, intermediate, and immediate criteria or between maximal and typical performance refers to types of criteria. Both organizational records and subjective evaluations (methods) can be used to assess them.

Organizational records can be further classified into direct measures of productivity and personnel data (Schmidt, 1980). Direct measures of productivity stress the number of units produced. Also included are measures of quality such as the number of errors, scrap material produced, etc. Personnel data, on the other hand, do not directly measure productivity but inferences of productivity can be derived based on them. Lateness or tardiness, tenure, absences, accidents, promotion rates, and filing grievances can be considered as indirect measures of productivity – there is an inferential leap involved in using these personnel data as measures of individual job performance. Organizational records, by focusing on observable, countable, discrete outcomes, may overcome the biasing influences of subjective evaluations but may be affected by criterion contamination and criterion deficiency just like subjective evaluations. Contamination occurs in that outcomes could be due to factors beyond the control of the individuals; deficiency occurs because the outcomes assessed may not take into account important aspects of individual job performance.

Subjective evaluations can be either ratings or rankings of performance. Ratings are criterion-referenced judgments where an individual is evaluated without reference to other individuals. The Graphic Rating Scale (GRS) is commonly used and several different formats have been introduced. The formats differ in the number of scale points, the clarity or discreteness of the scale points, etc. Empirical research suggests that psychometric properties are not affected by issues such as the number of scale points (Austin & Villanova,

1992). However, providing a common frame of reference across raters as to what each scale point refers to (e.g., what does an evaluation of 2 on a scale of 1–5 mean in behavioral terms?) has been suggested as an aid to improving consistency across and within raters. In fact, the behaviorally anchored rating scales (BARS) were developed based on this logic. To address the concern that some raters will be more comfortable recording observed behavior rather recording their evaluations of them, procedures such as checklists, weighted checklists, and behavioral observation scales (BOS) have been introduced. However, in a seminal review, Landy and Farr (1980) found that the different rating-scale formats do not make a large difference in the quality of the ratings.

Some attempts have been made to address the problem that raters could intentionally distort their ratings (especially the issue of rater leniency), by designing scales where the raters are not sure of the scoring rules. Forced Choice Scales and Mixed Standard Scales (MSS) are two such attempts. In a Forced Choice assessment, raters are provided with two equally favorable statements, only one of which discriminates between good and poor performers. The idea is that the rater who wants to give lenient ratings may choose the favorable but nondiscriminating statement as descriptive of the ratee. The MSS comprises three statements for each dimension of performance rated, with the three statements depicting an excellent, average, and poor performance, respectively, on that dimension. The rater rates the performance of each ratee as better than, equal to, or worse than the performance depicted in that statement. Statements across dimensions are mixed. The objective is to check whether a rater who rates an employee behavior as better than an excellent statement also provides better-than ratings for statements depicting average or poor performance in that dimension (Blanz & Ghiselli, 1972). Although such scales could reduce leniency and identify careless or inconsistent raters, their acceptability by raters has been found to be low (Austin & Villanova, 1992; Landy & Farr, 1980).

In contrast to ratings which are criterion-referenced assessments, rankings are norm-referenced assessments. The simplest form of ranking is to rank all ratees from best to worst. The ranking will depend on the set of ratees and it is impossible to compare the rankings from two different sets of individuals. The worst in one set may be better than the best in the second set of ratees. A modified version, called alternate ranking, involves 1) picking the best and worst ratees in the set of ratees under consideration, 2) removing the two chosen ratees, 3) picking the next best and worst from the remaining ratees, and 4) repeating the process until all ratees are ranked. The advantage of the alternate ranking method is that it reduces the cognitive load on the raters. Yet another approach is to compare each ratee to every other ratee, a method of paired comparisons that becomes unwieldy when the number of ratees increases. Finally, forced distribution methods can be used where a fixed percentage of ratees are placed in each level. Forced distribution methods can be useful to generate the desired distribution (mostly normal) of assessed scores. However, it is an open question whether such distributions reflect reality.

Sources of Ratings

With subjective evaluations (ratings or rankings), the question of who should rate arises. Typically, in traditional organizations the supervisors of the employees provide the ratings. Recent years have seen an increase in the use of 360-degree feedback systems (Church &

Bracken, 1997) where rating assessments can be made by the ratee himself or herself (self), subordinates, peers, and customers or clients. However, when ratings are used for administrative and personnel-selection-related purposes, self-ratings are mostly inappropriate. Traditionally, supervisory ratings have been used for validating selection predictors, to make promotion and selection decisions, etc. For example, Lent, Aurbach, and Levin (1971) found in their review of 1,506 validation studies that 63% of the studies used ratings as the criterion measurement method. Of these studies, 93% used supervisory ratings. Bernardin and Beatty (1984) estimated that over 90% of the ratings used in the literature are supervisory evaluations.

In addition to the use of supervisory ratings, peer ratings are also used in validation research. Lent et al. (1971) reported that the remaining 7% (after the 93% that used supervisory ratings) of validation studies used peer ratings to measure the criterion. Given that the traditional hierarchical structure of organizations is being replaced by more team-based work (Norman & Zawacki, 1991), the use of peer ratings in personnel selection research and practice is likely to increase. Several researchers have argued that the validity of predictors may differ depending on the use of peer or supervisor ratings. For example, Conway and Huffcutt (1997) suggest that the validity of predictors differs based on the source of the ratings. This is basically a hypothesis of differential validity by job performance rating source.

Some boundary conditions on our discussion of differential validity here are to be noted. Differential validity has also been claimed based on the content of the criterion. For example, it has been argued that personality will better predict teamwork and ability will better predict productivity. We will take up this form of differential validity based on content in a subsequent section in this chapter. Further, the term differential validity has also been used to assess whether predictor–criterion combinations are the same for different groups of individuals (e.g., Whites, Blacks). In this chapter we do not discuss differential validity in terms of different validity coefficients for the same predictor–criterion combinations for different groups of individuals. In this section, we are referring only to differential validity based on peer versus supervisory ratings.

When we discuss the potential for differential validity based on source of ratings (peer or supervisors), we are essentially discussing the equivalence of the two sources of ratings. This equivalence can be assessed either by estimating their intercorrelation or by assessing the pattern of correlations the two sources of ratings have with external variables. The first line of evidence focuses on the internal structure of the construct assessed by the two sources, whereas the second line of evidence explores the cross-structure of measures of job performance with measures of other constructs (Nunnally & Bernstein, 1994).

Several theoretical mechanisms have been proposed as to why peers and supervisors should differ. Most prominently, opportunity to observe has been postulated to differ across the two sources. In addition, Borman (1974) suggests that the objectives may be different for peers and supervisors. However, empirical evidence is not especially supportive of these proposed mechanisms. For example, Albrecht, Glaser, and Marks (1964) found that the convergence between peer and supervisor ratings in rating sales ability was .74. Harris and Schaubroeck (1988) reported that the correlation between peer and supervisor ratings of overall job performance was .62. Such high values of overlap between the two sources of ratings suggest that the prospects of differential validity are remote. Further, it should be

noted that the reported value of .74 is uncorrected for measurement error and the value of .62 was based on an attenuation correction value (reliability) of .60. Recent research (Rothstein, 1990; Salgado & Moscoso, 1996; Viswesvaran et al., 1996) suggests that inter-rater reliability of supervisory ratings is .52 and that of peer ratings is .42, values which suggest that the convergence is much higher than the reported values of .74 and .62.

An additional point needs to be taken into account in judging the convergence value of .74 as reported above. Viswesvaran et al. (2002) make a distinction between construct-level convergence and rating difficulty. Viswesvaran et al. (2002) state:

> Agreement between raters can be reduced by the absence of agreement on the nature of the construct to be rated or by difficulty of rating a particular agreed upon dimension, or by both. The correlation between peer and supervisory ratings may be reduced because peers and supervisors are rating different constructs or perceived dimensions of job performance (i.e., lack of construct-level convergence) because of differences in their understanding of the exact nature of the dimensions. That is, they are actually rating somewhat different performance dimensions. Conversely, even when peers and supervisors are rating the same performance dimension (or construct), the correlation between peer and supervisor ratings of a performance dimension may be lower for one dimension than another because it is difficult to rate reliably, leading to lower supervisor-peer correlations . . . In this paper we refer to this effect or process as "rating difficulty" for the sake of brevity. The two effects, lack of construct-level convergence and rating difficulty, are confounded in the observed correlation between peer and supervisor ratings. (p. 346)

In assessing differential validity, the focus should be on construct-level disagreements. Just because two measures differ in their reliabilities, their external correlates may be different. Such differences do not constitute evidence of differential validity. In fact, by introducing measurement error into measures one can generate evidence of differential validity. As such, in this section we will review the correlation between peer and supervisor ratings after correcting for rating difficulty and measurement error. We will not consider in this section evidence of differential validity of peer and supervisor ratings based on observed correlations with external variables.

A key question essentially then becomes whether, for any given dimension of job performance, peers and supervisors are rating the same construct or performance dimension. If the answer is in the affirmative, the true score correlation between peer and supervisor ratings is expected to be 1.00 (within sampling error). The observed peer–supervisor correlation can be corrected for measurement error to determine whether corrected values are within sampling error of 1.0. To this end, the confidence intervals around the observed correlation are corrected for measurement error. The end points of the confidence intervals can be corrected with the same attenuation formula as the observed correlation (Hunter & Schmidt, 1990). If inter-peer and inter-supervisor reliability values are used to make the attenuation corrections, then what is unique or idiosyncratic to a particular supervisor or peer (not shared with other supervisors or peers, respectively) is considered to be measurement error. That is, the construct underlying peer ratings is defined as what is common across peers, and the construct underlying supervisor ratings is defined as what is common across supervisors.

Viswesvaran et al. (2002) reported, in a meta-analytic cumulation of the existing literature reporting peer–supervisor ratings correlations, that the overlap between the two

sources is substantial. For two-thirds of the dimensions they investigated there was construct-level convergence. Peers and supervisors were rating the same construct and observed correlations between the two sources of ratings were lowered primarily due to measurement error (i.e., disagreements and idiosyncrasies between peers or between supervisors). This conclusion that peers and supervisors are rating the same constructs is also borne out by several other large-scale studies.

Mount, Judge, Scullen, Sytsma, and Hezlett (1998) found in a large-scale study of performance ratings that a model which postulated separate latent factors for rater-level (supervisors, peers) did not do better than one where each rater was treated as an independent method. That is, there were more disagreements across raters belonging to the same level than there was shared variance across raters of the same level. A similar finding was reported by Facteau and Craig (2001) who used item response theory (IRT) and confirmatory factor analyses (CFA) to demonstrate the equivalence of peer and supervisor ratings. In short, ample evidence exists that peers and supervisors converge in their evaluations of the same job performance dimensions. In other words, they are simply different, randomly parallel methods for assessing the same sets of constructs.

How does this square with arguments that peers and supervisors emphasize, observe, and value different behaviors? A reference to the domain-sampling model of reliability (Nunnally & Bernstein, 1994) will be informative. In test construction, we have a domain of interest and there are several items that could be used to assess the construct defining that domain. Similarly, the different behaviors observed by peers and supervisors have specific variance associated with them (i.e., the item specific variance), but that specificity does not affect the construct of interest. Teamwork is teamwork is teamwork – whether measured by behaviors considered relevant by peers or by behaviors considered relevant by supervisors. This explanation is also compatible with the findings that the same individual differences variable (e.g., general mental ability, conscientiousness) is predictive of different workplace behaviors of interest. A conscientious individual who is likely to engage in behaviors that result in better ratings from peers is just as likely to engage in behaviors that will result in similar ratings from supervisors. Different behaviors and manifestations of underlying traits may be observed by peers and supervisors, but the construct domain sampled and assessed remains the same.

As a practical consequence, for assessment of job performance in personnel selection, we would recommend the collection of ratings from different sources (i.e., peers and supervisors), not because there is likelihood of differential validity but because of a more comprehensive sampling of the domain of performance. A composite based on both peer and supervisory ratings will result in a more reliable and valid assessment. Furthermore, user acceptability may be enhanced by using the multiple sources in validation.

The Construct Domain of Individual Job Performance

What is included in the construct domain of individual job performance? Essentially this question addresses what dimensions are part of the construct. There is no one correct set of dimensions, since just as a pie can be sliced in different ways, a construct can be sliced

into different sub-dimensions and facets that vary in terms of behavioral specificity, depending on the objectives of the researcher/practitioner. The attempt here is more to review the different dimensions or aspects of performance so as to glean an idea of what the construct of job performance entails. Further, given the numerous dimensions of job performance postulated in the extant literature, it might be confusing for a practitioner to select a subset of dimensions to assess. Defining the job performance construct domain for any job can, to a large extent, be guided by job-analytic data. However, it is also useful to recognize that similar categories of behaviors span across jobs and a summary of these main dimensions would be useful. For this purpose, we provide in Table 16.1 a summary of major job performance dimensions that have been discussed and utilized in the extant job performance literature.

Several strategies can be used to assess the dimensionality of the job performance construct. These include rational, theoretical, and factor analytic approaches. First, researchers have reviewed job performance measures used in different contexts and attempted to synthesize what dimensions make the construct of job performance. This rational method of synthesizing and theory building is, however, affected by the personal biases of the individual researchers. It is true that the factor analytic approach reviewed below is also influenced by the personal biases of researchers in the interpretation of the ensuing factor analytic results. However, compared to rational synthesis, there is an additional safeguard in factor analytic approaches, in that personal biases are checked by the empirical data collected and analyzed. Further, the cognitive load in integrating the vast number of dimensions proposed in the literature results cannot be denied. The same label has been used to refer to different dimensions as well as different labels for the same dimension (teamwork, interpersonal facilitation may overlap in many studies). In job performance assessments this has resulted in what personality psychologists have described as a jingle-jangle fallacy in personality assessments.

Second, researchers (e.g., Welbourne, Johnson, & Erez, 1998) have invoked organizational theories to define what the content of the job performance construct should be. Welbourne et al. used role theory and identity theory to explicate the construct of job performance. Borman and Motowidlo (1993) used the literature on socio-technical systems to specify that job performance should have two components: task and contextual performance that parallels the social and technical systems that are postulated to make up the organization.

Rational synthesis and theory-based specifications have to be empirically tested and factor analysis has been used to study the construct domain of job performance. In such an empirical approach, several measures of job performance are obtained from a sample of employees and their interrelationships assessed (e.g., Rush, 1953). The use of confirmatory factor analysis has enabled researchers to combine rational synthesis and empirical partitioning of variance. In a typical factor analytic study, individuals are assessed on multiple indices of job performance. Correlations are obtained between the measures of job performance and factor analysis is used to identify the measures that cluster together. Based on the commonalties across the measures that cluster together, a dimension is defined. For example, when absence measures, lateness measures, and tenure cluster together, a dimension of withdrawal behaviors is hypothesized.

The literature on the number of dimensions necessary to represent the domain has been contradictory. Rush (1953) factor analyzed nine rating measures and three

TABLE 16.1 Common job performance dimensions

Job performance dimension	*Description*
Productivity or task performance	This dimension typically refers to the actual counts of the units produced or ratings of the same, as well as ratings of behaviors deemed to constitute the core tasks of jobs.
Interpersonal competence	This refers to how well an individual behaves interpersonally at work as well as builds and maintains relationships in the work environment; can variously include teamwork, facilitating peers performance, etc.
Leadership	Behaviors associated with inspiring others, taking charge of situations for groups, bringing out extra performance in others, motivating others to scale great heights. Sometimes specific components such as leadership judgment and decision making could be stressed.
Effort	The persistence and initiative shown by individuals in getting tasks done. Sometimes lack of effort is reflected in facets of the counterproductive behavior dimensions such as tardiness, absences.
Job knowledge	Declarative and procedural knowledge to perform the job, including explicit and implicit rules and procedures to follow.
Counterproductive behaviors	Negative behaviors that detract from the value of employees to the organization, that are disruptive as they disrupt work-related activities, that are antisocial as they violate social norms, and that are deviant as they diverge from organizationally desired behaviors. Includes withdrawal behaviors, rule breaking, theft, violence, substance abuse on the job, sabotage, etc. Originally conceptualized as the polar opposite of citizenship behavior, recent empirical findings indicate that this is a separate dimension from citizenship behaviors.
Citizenship behaviors	Also referred to as contextual performance, prosocial behavior, altruism, etc. Refers to the extent an individual contributes to the welfare of the organization in ways not formally stated in job descriptions.

organizational-records-based measures of job performance for 100 salespeople and identified four factors: objective achievement, learning aptitude, general reputation, and proficiency of sales techniques. Baier and Dugan (1957) obtained data on 346 sales agents on fifteen objective variables and two subjective ratings and factor analysis of the 17 × 17 intercorrelation matrix resulted in one general factor. In contrast, Prien and Kult (1968) factor analyzed a set of 23 job performance measures and found evidence for seven distinct dimensions. Roach and Wherry (1970), using a large sample of ($N = 900$) salespersons, found evidence for a general factor whereas Seashore, Indik, and Georgopolous (1960), using comparably large samples ($N = 975$), found no evidence for a general factor.

Ronan (1963) conducted a factor analysis of a set of eleven job performance measures and found evidence for four factors. Gunderson and Ryman (1971) examined the factor structure of individual job performance in extremely isolated groups and suggested three dimensions: task efficiency, emotional stability, and interpersonal relations. Klimoski and London (1974) used multi-source data and reported evidence for the presence of a general factor, a finding that is interesting when considered in the wake of arguments that raters at different levels of job performance construe the content domain of job performance differently. Factor analytic studies in the last two decades (1980–99) have used much larger samples and refined techniques of factor analysis. However, each of these studies, even when they postulate the same number of dimensions, comes up with different dimensions. The four dimensions identified by Rush (1953) are not the same four dimensions presented by Murphy (1989). Sometimes, different names are used to refer to the same dimension whereas at other times the same label is used to refer to different dimensions.

Viswesvaran and Ones (2000) developed a two-dimensional grid to group and provide a format structure to these different taxonomies. The first dimension is whether the taxonomy was developed for a single occupation or is applicable across occupations. For example, Hunt (1996) developed a model of generic work behavior applicable to entry-level jobs especially in the service industry. Using performance data from over 18,000 employees primarily from the retail sector, Hunt (1996) identified nine dimensions of job performance that do not depend on job-specific knowledge. The nine dimensions were: adherence to confrontational rules, industriousness, thoroughness, schedule flexibility, attendance, off-task behavior, unruliness, theft, and drug misuse. Alternately, taxonomies can be developed that will be applicable across occupations. One such example is the eight-dimensional taxonomy provided by Campbell (1990), who describes the latent structure of job performance in terms of eight dimensions: job-specific task proficiency, non-job-specific task proficiency, written and oral communication, demonstrating effort, maintaining personal discipline, facilitating peer and team performance, supervision, and management or administration. The description of these eight dimensions is further elaborated in Campbell (1990) and Campbell, McCloy, Oppler, and Sager (1993). Five of the eight dimensions were found in a sample of military jobs (Campbell, McHenry, & Wise, 1990).

The second dimension that Viswesvaran and Ones (2000) used to group the different taxonomies is the focus on specific performance aspects versus clusters of performance aspects. For example, Smith, Organ, and Near (1983) popularized the concept of "organizational citizenship behavior" (OCB) in the job performance literature. Recently, taxonomies of counterproductive behaviors have been proposed (Gruys & Sackett, 2003). Here the focus is on certain aspects of performance and not on the overall job

performance construct. The goal is not to define the entire construct domain of job performance but to home in on specific sub-dimensions.

This empirical approach to specifying the construct of job performance is limited by the number and type of measures included in the data collection phase. Recently, the combination of meta-analysis and structural equations modeling (Viswesvaran & Ones, 1995) has greatly extended this approach. No longer are we limited to the number of measures that can be administered to one sample of employees. As long as we can estimate the correlation between different measures (even based on different samples), a structural equations modeling of the meta-analyzed correlation matrix can be employed to investigate the factor structure of job performance.

Viswesvaran (1993) combined meta-analyses and structural equations modeling to investigate the factor structure of job performance. A large general factor was found across the different measures. To avoid biases in judgmentally describing the construct domain of job performance, Viswesvaran (1993) invoked the lexical hypothesis from personality literature (Goldberg, 1995). The lexical hypothesis states that practically significant individual differences in personality are encoded in the language used, and therefore, a comprehensive description of personality can be obtained by collating all the adjectives found in the dictionary. Extending this principle to job performance assessment suggests that a comprehensive specification of the content domain of the job performance construct can be obtained by collating all the measures of job performance that have been used in the extant literature.

The model of job performance that emerges from the meta-analytic cumulation by Viswesvaran (1993) views the various measures of job performance, such as quality and quantity of work performance, absenteeism, turnover, violence on the job, and teamwork, as the manifestations of a general construct of job performance. This can be stated in factor analytic terms as follows. The standing of an individual on any specific performance measure (e.g., absenteeism) can be *hypothesized* to depend on the general factor (i.e., overall job performance), the group factor (e.g., the withdrawal behavior of the employee; absenteeism, tardiness, time theft, turnover all may belong to this group), the specific factor (i.e., absenteeism), and a random error component. There may or may not be group factors for a specific measure of job performance such that that measure of job performance correlates more with the measures of job performance in that group than with the measures of job performance in any other group. The existence of such clusters or groups of measures of job performance is an empirical question. Demonstrating the existence of such a hierarchy of performance measures is an empirical question that depends on an investigation of the true score intercorrelations between the different measures of job performance. Viswesvaran (1993), on meta-analyzing over 2,600 correlations, concluded that a general factor exists across all measures of job performance used in the extant literature over the past 100 years.

Recently, Viswesvaran, Schmidt, and Ones (in press) refined this analysis to disentangle the effects of idiosyncratic halo error from this general factor. Cumulating results across over 300 studies, Viswesvaran et al. (in press) estimated the true score correlations across different dimensions of job performance. Within-peer and within-supervisor correlations were analyzed separately from between-supervisor and -peer correlations. The within or intra-rater correlations (same peer or same supervisor rates both dimensions being corre-

lated) are affected by halo and measurement error whereas the between-rater correlation (peers rating one dimension and supervisors rating the other in any inter-dimension correlation) is not affected by halo. Similarly, within-rater reliability (i.e., alphas) is inflated by halo but not interrater reliabilities (which are actually lowered by halo). Thus correcting within-rater correlations with within-rater reliabilities accounts for measurement error but not halo. Correcting interrater correlations (peer–supervisor) with interrater reliabilities corrects for both halo and measurement error. A comparison of these two sets of correlations estimates the inflationary effects of halo. Viswesvaran et al. (in press) present evidence to suggest the presence of a general factor (that explains 27% to 54% of variance) across job performance dimensions, even after accounting for rater idiosyncratic halo. Cumulative empirical evidence clearly supports the presence of a general factor and hence a hierarchical view where specific dimensions of job performance all load onto a higher-order factor in varying degrees.

The presence of a general factor raises the question of whether differential validities will be found in using the same predictor for different dimensions of performance. A job performance measure, when used as a criterion in a validation study, is a standard used for two purposes (Schmidt, 1980): a) to decide which selection procedures to use and which to reject (i.e., to determine the best test battery); and b) to weight the selection procedures selected for use. As such, any two job performance measures are equivalent if their use leads to the adoption of the same selection procedures and assignment to them of the same relative weights. Schmidt (1980) reports that in one large sample study done in the army it was found that a job sample criterion, supervisory ratings, and a job knowledge measure all resulted in the adoption of the same selection procedures and assignment of essentially identical relative weights. Oppler, Sager, McCloy, and Rosse (1993) found, using a sample of 3,086 soldiers, that prediction composites developed using job knowledge tests as the criterion compared favorably in validity to those developed using hands-on tests. Nathan and Alexander (1988) investigated whether or not job performance measurement method moderates validities of cognitive ability tests. They meta-analytically cumulated the validities reported for supervisory ratings, rankings, work samples, production quantity, and production quality. No evidence was found for differential validity.

Despite these results, researchers have continued to focus on the question of differential prediction. In examining differential prediction, it is essential to keep in mind the influence of sampling error, which greatly influences the results when multiple regression strategies are used to develop batteries that include correlated predictors (Hunter, Crosson, & Friedman, 1985). Multiple regression does not work very well (especially with correlated predictors, which inflate sampling error in the regression weight estimates) unless sample sizes are extremely large (Helme, Gibson, & Brogden, 1957).

Recent years have seen theoretically based tests of differential validity. Borman and Motowidlo (1993) postulated that ability will predict task performance more strongly than individual differences in personality. On the other hand, individual differences in personality were hypothesized to predict contextual performance better than ability. However, empirical evidence is not wholly supportive of this claim. Alonso (2001) cumulated the literature on a) personality predicting contextual and task performance, and b) ability predicting contextual and task performance. Across 512 validity coefficients, Alonso found that cognitive ability predicted both task and contextual performance. Some personality

variables were found to be predictive of contextual performance. This lack of strong empirical support for differential validity based on the content of the criterion is perhaps due to the general factor in job performance assessment. Just as we did not find differential validity for source of ratings (peers and supervisors), there is no evidence of differential validity based on content of the criteria.

In personnel selection, writers interested in downplaying the importance of traits such as ability (where there are large ethnic group differences) have sometimes pinned their hopes on findings of differential validity by what is included in the content of the criterion. The hope was that individuals high on ability may score high on some dimensions of job performance but not on others. If true, every employee could be in the top 10%, albeit in different dimensions of performance (evaluated by different sources). Cumulative empirical evidence, however, is not supportive of this Polyanna-ish view of success. The prospects of differential validity in personnel selection based either on the dimensional content or on the source of assessment are improbable (see also Viswesvaran et al., in press).

Emerging Issues

In this section, we address four issues: 1) definitional issues surrounding the job performance construct, 2) issues in assessing the reliability of job performance assessment, 3) job performance assessment for personnel selection in a team context, and 4) job performance measurement in an international context for expatriate selection and assignment. We should note that for some of the topics to be discussed the empirical data is scant or virtually non-existent. When confronted with such situations, we raise the relevant issues for future research to consider.

Definitional issues

We noted that individual job performance refers to behaviors that can be evaluated. Although we used the term behaviors, we note that the difference between behaviors and outcomes is not clear cut in many instances. Some researchers (Campbell, 1990) insist on a clear demarcation between behaviors and outcomes. The main thrust of this argument is that individuals should be evaluated on what they can control. Other researchers (Austin & Villanova, 1992; Bernardin & Pence, 1980) de-emphasize this difference between behaviors and outcomes. However, in many instances it is not clear what is under the control of an employee. Consider the research productivity of a professor. A relevant behavior to evaluate in the context of this job is writing research papers. But what gives meaning to such a behavior are factors such as whether the papers written are published and, if so, the quality of the outlets. The number of papers published is certainly influenced by many factors beyond the control of the professor. Even the number of papers *written* is influenced by factors outside the control of a professor. Thus, this distinction between behaviors and outcomes is something to be evaluated in the choice of a measure for assessing job performance in personnel selection.

Second, researchers need to pay more attention to temporal relations between different dimensions of job performance. Is teamwork likely to increase productivity? That is, instead of studying predictor–criterion relationships of the type x → y, researchers need to investigate relationships of the type x → y1 → y2 (see Alonso, Viswesvaran, & Sanchez, 2001, for an illustrative example). To some extent this has been explored in team and group dynamics literature where some dimensions are construed as process variables. At this point we want to be clear that calling for an investigation of dynamic relations between job performance dimensions is different from the issue of criterion dynamicity. Criterion dynamicity refers to whether the measurement of a particular dimension changes over time. The changes could be either in the mean performance levels or in the relationships with other variables (of which test–retest reliability is a special case of relationship with same variable). Cumulative evidence (Barrett, Caldwell, & Alexander, 1989) indicates that job performance measures are stable over time.

Reliability issues in job performance assessment

The question of the appropriate reliability coefficient to use in personnel selection validation studies has also occupied IWO psychologists in recent years (Murphy & DeShon, 2000; Schmidt, Viswesvaran, & Ones, 2000). Of course, one has to answer a more fundamental question of whether any corrections are to be made to observed validity coefficients. Arguments have been made (Outtz, 2002) that corrections distort "what is" from "what could be" (see DeShon, 2002, for a harsh but misguided critique of reliability corrections). Another argument is that researchers should be conservative in their estimates – thus, although unreliability lowers the validity coefficients, practitioners should take the conservative value in evaluating the success and utility of their predictors.

It is the corrected correlations that reflect what is (Viswesvaran, 2003). The observed uncorrected (note, we are discussing criterion unreliability here) correlations reflect merely what researchers were able to do with their criterion measurement and the quality of their data collection and does not reflect the validity or utility of the selection tool (see Sackett, Schmidt, Tenopyr, & Kehoe, 1985; Schmidt, Hunter, Pearlman, & Rothstein, 1985). Finally, the objective in science should be to obtain precise estimates and not conservative estimates. Even in a practical application, the question of conservative in what direction arises – similar to the Type I and Type II errors in statistical tests. Thus, correcting for criterion unreliability in personnel selection validation studies is always an essential step (Viswesvaran et al., 1996).

Given that correcting for unreliability in the criterion is essential, the question becomes that of what reliability coefficient should be used. Different reliability coefficients treat different sources of variance as error (Guion, 1998; Schmidt & Hunter, 1996). Broadly speaking, there are three types of variance – transient error, rater idiosyncratic variance, item-specific variance when several items are used to measure a dimension, and random response error. All reliability coefficients take into account random response error. The coefficient alpha that is used in most studies (Cortina, 1993) treats item-specific variance and random response as errors.

However, in personnel selection validation research we want to generalize our findings across raters. There really is no logic in claiming that our predictor scores are likely to

predict job performance as measured by one idiosyncratic rater. Rater idiosyncratic variance is a large component in the variance of job performance ratings; Viswesvaran et al. (in press) estimate this component to be as much as 30%. Only interrater reliability estimates this error component and, as such, interrater reliabilities are the only appropriate reliability coefficient (cf. Schmidt et al., 2000) in personnel selection validation research (and in research that attempts to generalize findings across raters).

Thus, the choice of reliability coefficient used to assess measurement error in criteria has enormous practical implications. The hallmark of a good predictor in personnel selection is its high criterion-related validity. Criterion-related validity is a correlation between predictor scores and criterion scores. Given that most of the criterion data are obtained from ratings (cf., Bernardin & Beatty, 1984; Viswesvaran et al., 1996), failure to account for the idiosyncrasies of an individual rater distorts our validation efforts. Pragmatic science (Anderson, Chapter 1, this volume; Anderson, Herriot, & Hodgkinson, 2001) requires the use of an appropriate reliability coefficient – in this instance, interrater reliability.

Of course, the realities of assessing interrater reliability should also be considered. In an organization, the same pair of supervisors does not assess all employees. Thus, different pairs of supervisors will be rating different individuals, and to estimate interrater reliability we will arbitrarily designate one rater as Rater 1 and the other as Rater 2. That is, Rater 1 and Rater 2 may be different raters for different individuals; the only restriction is that Rater 1 and Rater 2 should be different for the same individual employee. Some researchers (Murphy & DeShon, 2000) have focused on this fact to argue that interrater correlations do not estimate interrater reliability.

It is important to note that the realities of data collection outlined above merely include a new component into the reliability estimation. This is the rater main effect or leniency/harshness. While research must investigate factors that influence this variance component, in personnel selection practice it is important to correct for rater idiosyncrasies. We do not want to design a selection system to predict job performance as idiosyncratically defined by a single rater. Our predictors should predict performance as defined consensually. Legal and fairness concerns demand such professional practice. Perhaps recourse to the concept of "natural distance" that Anderson (this volume) advocates will be useful here. While the best practice of using interrater reliability is used, future research should empirically examine the influence of rater main effects on outcomes.

The issues involved can be further elucidated within the framework that Anderson (Chapter 1, this volume) advances. Anderson notes four scenarios of potential interaction between the science and practice of personnel selection. One of the scenarios involves unreliable findings influencing practice. In assessing the reliability of job performance measures, arguments were made that even if the same two supervisors rate all employees, intercorrelation does not estimate reliability because the two raters are not (strictly) parallel. For example, raters differ in leniency. However, later research showed that strict parallelism is not required and raters are randomly equivalent (the difference in reliability estimates between the assumption of random equivalence and strict parallelism was only .02). Thus, a lot of concern was expressed over a trivial issue. Consider another example. For a long time it was argued that raters at different levels (i.e., positions) observe different behaviors. Thus, even if the same two supervisors rate all employees, they will have different perspectives due to their different role relationships with the employee, and there-

fore, the intercorrelation between their ratings fails to assess reliability. However, as our review of peer–supervisor convergence showed, despite such widespread claims there was no construct-level disagreement between peers and supervisors. This is another example of unreliable findings influencing practice. To avoid similar mishaps in future, research should evaluate different assumptions and sources of variance in ratings at a safe distance (natural distance?) while practitioners obtain the best estimate possible for the criterion with interrater reliability.

A related issue has been raised by some researchers (e.g., Morris & Lobsenz, 2003; Murphy & DeShon, 2000). The argument is that classical measurement theory is limited and that generalizability coefficients are appropriate. This argument is logically flawed. Both classical measurement theory and generalizability theory can be used to assess the different sources of error **if the appropriate data are collected**. If researchers want to estimate the generalizability coefficient where rater idiosyncratic variance is construed as error, they have to collect ratings from two raters. Similarly, if researchers also want to generalize over time, the raters should provide ratings at two different points in time. But if such data are available, classical methods of reliability estimation can be profitably employed. Researchers merely need to correlate the ratings given by one rater at one time with the ratings given by the second rater at an alternate time. These issues are pertinent to the arguments presented by Murphy and DeShon (2000) and Murphy (2003). The authors argue initially about how classical reliability estimates are not appropriate for correcting validity coefficients and conclude their analyses by recommending the use of generalizability coefficients, although all the sins they visit on interrater correlation is also applicable to generalizability coefficients (and if different data are available in the generalizability assessments, the appropriate interrater correlations can be easily computed). Thus, it is erroneous to claim that generalizability coefficients provide more information than classical reliability estimates. Both can yield the same information, provided the appropriate data are collected and analyzed.

Assessment of team performance

Teams are widely used in organizations (Sundstrom, DeMeuse, & Futrell, 1990). The increasing complexities of work and technological advances have necessitated, and at times facilitated, the use of teams. The composition of individual performance to assess team and group performance is an important area. In fact, volumes have been written on assessment of team performance (cf., Swezey & Salas, 1992). Our purpose here is not to review issues in team performance assessments but to show how developments in that field influence job performance assessment in personnel selection settings.

Consider our delineation of what a criterion should do for personnel selection (see also Schmidt, 1980). In personnel selection, the job performance assessed is used to validate predictors. Once validated, the predictors are used to select employees from a pool of applicants. Viewed from this functional perspective, a large number of unanswered questions about assessments of team performance arise. We summarize some of the questions in Table 16.2.

First, consider the definition of a team. A broad definition suggests that teams involve two or more people who interact dynamically and share a common goal (Reilly &

TABLE 16.2 Team performance: issues to consider in personnel selection

Identifying level of team aggregation
Identifying dimensions of job performance that are common across levels of aggregation
Identifying dimensions of job performance that are unique to one level
Specifying composition models
Assessing equivalence of rater techniques and methods across levels
Assessing equivalence of rater cognitive processes/biases in individual and team evaluations
Distinguishing between assessing individual performance of employees in teams from team performance

McGourty, 1998; Salas, Dickinson, Converse, & Tannenbaum, 1992). However, by this definition an entire organization can be considered as a team. In fact, one can extend this to say that a particular industry is a team. Extended further, we can say that an entire economy is a team. For personnel selection purposes, we need to be more specific about the definition of our team. We have to specify whether we are interested in selecting individuals to work as a defined task group, or as an organizational member, or assessing candidates for their fit to occupations (e.g., vocational counseling). Stevens and Campion (1994) proposed a predictor to select individuals for work teams, although the discriminant validity of the knowledge measure from individual cognitive ability was not robust. Person–organization fit measures have been proposed, although not used much in selection. Interest inventories have been proposed to assess suitability for occupations (again not used widely in selection contexts).

Once we have decided whether we are selecting an individual or an individual for a team (level specified most likely to be groups), other questions arise. First, are the performance dimensions identified at the individual level applicable to the group level? Are there new dimensions that emerge at team level? What are the implications of some of these dimensions (e.g., cohesion) for selecting individuals? What are the individual differences variables that relate to these dimensions of team performance? Although the techniques used to assess individual performance (records, ratings, etc.) could be helpful in assessing team performance, sources of ratings may differ. In evaluating performance in teams, there could be a greater emphasis on peer assessments, for example.

Assessments of performance in international contexts

As noted earlier, increasing globalization is a fact of life (Anderson et al., 2001) and the science and practice of personnel selection have to accommodate this fact. There is a large literature (Sinangil & Ones, 2001) on expatriate selection, and assessment of individual job performance in international contexts is a critical issue. Some of the questions that arise in this context are summarized in Table 16.3.

First, are the existing dimensions (identified in Table 16.1) similar in international contexts? Are measurement techniques and rating scales comparable? In our review of rating scales used in personnel selection, we noted how graphic rating scales and behaviorally

TABLE 16.3 Issues in job performance assessment in international contexts

Are existing dimensions (cf., Table 16.1) the same in different cultures? Are there new dimensions of job performance when assessing performance in international contexts?
Are the behaviors associated with performance dimensions the same in different cultures?
Are measurement techniques/rating scales comparable across cultures?
Are the relative weights given to the different dimensions in assessing overall performance the same across cultures?
Which raters (i.e., rating sources) have face validity and are deemed acceptable in different cultures?
What dimensions should be used in validating predictors for expatriate selection?
What factors differentially influence the collection of performance appraisal data across cultures?

anchored rating scales are preferred by users more than mixed standard scales (where it is not clear to the rater what rating is being given). Will this result translate to cultures where users are more likely to accept larger power distances and in cultures that tolerate uncertainty?

Are there new dimensions that need to be included when considering individual job performance in a global context? Some research (e.g., Conner, 2000; Kanter, 1995) suggests that individuals should develop a global mindset to succeed in a globalized economy. In the area of expatriate selection, research stresses the need to assess individual performance on dimensions that relate to cultural adjustment (Caligiuri, 2000; Deshpande & Viswesvaran, 1992), although Ones and Viswesvaran (2001) have argued that adjustment is best considered a determinant of expatriate performance rather than a sub-dimension. The source of ratings deemed acceptable may differ based on organizational culture.

Conclusions

New dimensions of job performance are appearing. Given that the concept of job is changing, it is an open question whether we would be discussing task and work performance in future. Technological assessments have provided new tools to obtain measurements and sometimes provided new measures (electronic performance monitoring). In personnel selection assessment issues in measuring job performance may change over the decades, but the centrality of the construct of job performance is likely to remain undimmed. Scientists and practitioners should gain a comprehensive understanding of the issues in job performance measurement and assessment to be effective in personnel selection. Hopefully, this chapter has summarized the important findings to date so as to give readers an understanding of the importance of this construct to personnel selection.

Note

The order of authorship is arbitrary; both authors contributed equally.

References

Albrecht, P. A., Glaser, E. M., & Marks, J. (1964). Validation of a multiple-assessment procedure for managerial personnel. *Journal of Applied Psychology, 48*, 351–360.

Alonso, A. (2001). *The relationship between cognitive ability, big five, task and contextual performance: A meta-analysis.* Unpublished Master's Thesis, Florida International University, Miami, FL.

Alonso, A., Viswesvaran, C., & Sanchez, J. I. (2001, April). *Mediating roles of task and contextual performance on predictor validity: A meta-analysis.* Poster presented at the 16th annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA.

Anderson, N., Herriot, P., & Hodgkinson, G. P. (2001). The practitioner-researcher divide in Industrial, Work and Organizational (IWO) psychology: Where are we now, and where do we go from here? *Journal of Occupational and Organizational Psychology, 74*, 391–411.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77*, 836–874.

Baier, D. E., & Dugan, R. D. (1957). Factors in sales success. *Journal of Applied Psychology, 41*, 37–40.

Barber, A. E. (1998). *Recruiting employees: Individual and organizational perspectives.* Thousand Oaks, CA: Sage.

Barrett, G. V., Caldwell, M. S., & Alexander, R. A. (1989). The predictive stability of ability requirements for task performance: A critical reanalysis. *Human Performance, 2*, 167–181.

Bernardin, H. J., & Beatty, R. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent-PWS.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60–66.

Blanz, F., & Ghiselli, E. E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology, 25*, 185–199.

Blum, M. L., & Naylor, J. C. (1968). *Industrial psychology: Its theoretical and social foundations.* New York: Harper & Row.

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior and Human Performance, 12*, 105–124.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco, CA: Jossey Bass.

Brogden, H. E. (1946). An approach to the problem of differential prediction. *Psychometrika, 11*, 139–154.

Caligiuri, P. M. (2000). The Big Five personality characteristics as predictors of expatriate desire to terminate the assignment and supervisor-rated performance. *Personnel Psychology, 53*, 67–88.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, 2nd ed., pp. 687–731). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey-Bass.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43*, 313–333.

Church, A. H., & Bracken, D. W. (1997). Advancing the state of the art of 360 degree feedback. *Group and Organization Management, 22*, 149–161.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130–135.

Conner, J. (2000). Developing the global leaders of tomorrow. *Human Resource Management, 39* (2 & 3), 147–158.

Conway, J. M., & Huffcut, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331–360.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

DeShon, R. (2002). Generalizability theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 189–220). San Francisco: Jossey-Bass.

Deshpande, S. P., & Viswesvaran, C. (1992). Is cross-cultural training of expatriate managers effective? A meta-analysis. *International Journal of Intercultural Relations, 16*, 295–310.

DeVries, D. L., Morrison, A. M., Shullman, S. L., & Gerlach, M. L. (1986). *Performance appraisal on the line.* Greensboro, NC: Center for Creative Leadership.

DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and white–black differences. *Journal of Applied Psychology, 78*, 205–211.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86*, 215–227.

Goldberg, L. R. (1995). What the hell took so long? Donald Fiske and the big-five factor structure. In P. E. Shrout & S. T. Fiske (Eds.), *Advances in personality research, methods, and theory: A festschrift honoring Donald W. Fiske.* New York: Erlbaum.

Gruys, M. L., & Sackett, P. L. (2003). Investigating the dimensionality of counterproductive work behavior. *International Journal of Selection and Assessment, 11*, 30–42.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel selection.* Mahwah, NJ: Lawrence Erlbaum.

Gunderson, E. K. E., & Ryman, D. H. (1971). Convergent and discriminant validities of performance evaluations in extremely isolated groups. *Personnel Psychology, 24*, 715–724.

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

Helme, W. E., Gibson, N. A., & Brogden, H. E. (1957). *An empirical test of shrinkage problems in personnel classification research.* Personnel Board, Technical Research Note 84.

Hunt, S. T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology, 49*, 51–83.

Hunter, J. E., Crosson, J. J., & Friedman, D. H. (1985). *The validity of ASVAB for civilian and military job performance.* Technical Report.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting for error and bias in research findings.* Newbury Park, CA: Sage.

Kanter, R. M. (1995). *World class: Thinking locally in a global economy*. New York: Simon & Schuster.

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

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

Lent, R. H., Aurbach, H. A., & Levin, L. S. (1971). Predictors, criteria, and significant results. *Personnel Psychology, 24*, 519–533.

Mace, C. A. (1935). *Incentives: Some experimental studies.* (Report 72). London: Industrial Health Research Board.

Malos, S. B. (1998). Current legal issues in performance appraisal. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 49–94). San Francisco: Jossey-Bass.

Morris, S. B., & Lobsenz, R. (2003). Evaluating personnel selection systems. In J. E. Edwards, J. C. Scott, & N. S. Raju (Eds.), *The human resources program-evaluation handbook* (pp. 109–129). Thousand Oaks, CA: Sage.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology, 51*, 557–576.

Murphy, K. R. (1989). Dimensions of job performance. In R. Dillon & J. Pelligrino (Eds.), *Testing: Applied and theoretical perspectives* (pp. 218–247). New York: Praeger.

Murphy, K. R. (Ed.). (2003). *Validity generalization: A critical review*. Mahwah, NJ: Erlbaum.

Murphy, K. R., & DeShon, R. (2000). Inter-rater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.

Nathan, B. R., & Alexander, R. A. (1988). A comparison of criteria for test validation: A meta-analytical investigation. *Personnel Psychology, 41*, 517–535.

Norman, C. A., & Zawacki, R. A. (1991, December). Team appraisals – team approach. *Quality Digest, 11*, 68–75.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Ones, D. S., & Viswesvaran, C. (2001). Integrity tests and other Criterion-focused Occupational Personality Scales (COPS) used in personnel selection. *International Journal of Selection and Assessment, 9*, 31–39.

Oppler, S. H., Sager, C. E., McCloy, R. A., & Rosse, R. L. (1993, May). The role of performance determinants in the development of prediction equations. In F. L. Schmidt (Chair), *Job performance: Theories of determinants and factor structure*. Symposium conducted at the eighth annual meeting of the Society of Industrial and Organizational Psychologists, San Francisco.

Outtz, J. L. (2002). The role of cognitive ability tests in employment selection. *Human Performance, 15*, 161–171.

Prien, E. P., & Kult, M. (1968). Analysis of performance criteria and comparison of a priori and empirically-derived keys for a forced-choice scoring. *Personnel Psychology, 21*, 505–513.

Reilly, R. R., & McGourty, J. (1998). Performance appraisal in team settings. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 244–277). San Francisco: Jossey-Bass.

Roach, D. E., & Wherry, R. J. (1970). Performance dimensions of multi-line insurance agents. *Personnel Psychology, 23*, 239–250.

Ronan, W. W. (1963). A factor analysis of eleven job performance measures. *Personnel Psychology, 16*, 255–267.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75*, 322–327.

Rush, C. H. (1953). A factorial study of sales criteria. *Personnel Psychology, 6*, 9–24.

Sackett, P. R., Schmitt, N., Tenopyr, M. L., & Kehoe, J. (1985). Commentary on forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697–798.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482–486.

Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Towards an understanding of team performance and training. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 3–29). Norwood, NJ: Ablex.

Salgado, J. F., & Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor Skills, 83*, 1195–1201.

Schmidt, F. L. (1980). *The measurement of job performance*. Unpublished manuscript.

Schmidt, F. L., & Hunter, J. E. (1992). Causal modeling of processes determining job performance. *Current Directions in Psychological Science, 1*, 89–92.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Rothstein, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697–798.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901–912.

Seashore, S. E., Indik, B. P., & Georgopoulos, B. S. (1960). Relationships among criteria of job performance. *Journal of Applied Psychology, 44*, 195–202.

Sinangil, H. K., & Ones, D. S. (2001). Expatriate management. In N. Anderson, D. Ones, H. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, & organizational psychology: Vol. 1, Personnel psychology* (pp. 424–443). London: Sage.

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 68*, 655–663.

Smith, P. C. (1976). Behavior, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745–775). Chicago: Rand McNally.

Stevens, M. J., & Campion, M. A. (1994). The knowledge, skill, and ability requirements for teamwork: Implications for human resource management. *Journal of Management, 20*, 503–530.

Sundstrom, E., DeMeuse, K. P., & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist, 45*, 120–133.

Swezey, R. W., & Salas, E. (Eds.). (1992). *Teams: Their training and performance*. Norwood, NJ: Ablex.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

Viswesvaran, C. (1993). *Modeling job performance: Is there a general factor?* Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Viswesvaran, C. (2003). [Review of *Measuring and analyzing behavior in organizations*. San Francisco: Jossey-Bass, 2002, 591 pages.]. *Personnel Psychology, 56*, 283–286.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48*, 865–885.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment, 8*, 216–227.

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.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology, 87*, 345–354.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (in press). Is there a general factor in job performance ratings? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*.

Welbourne, T. M., Johnson, D. E., & Erez, A. (1998). The role-based performance scale: Validity analysis of a theory-based measure. *Academy of Management Journal, 41*, 540–555.

Whisler, T. L., & Harper, S. F. (Eds.). (1962). *Performance appraisal: Research and practice*. New York: Holt.