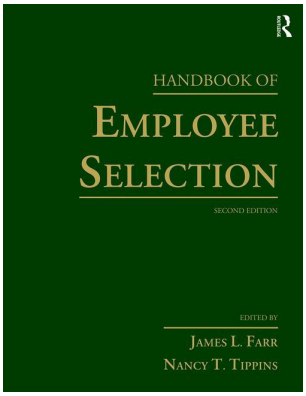

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

The Measurement of Task Performance as Criteria in Selection Research

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-20
Walter C. Borman, Matthew R. Grossman, Rebecca H. Bryant, Jay Dorio
Published online on: 22 Mar 2017

20

The Measurement of Task Performance as Criteria in Selection Research

WALTER C. BORMAN, MATTHEW R. GROSSMAN,
REBECCA H. BRYANT, AND JAY DORIO

This chapter is about measuring task performance (i.e., the technical proficiency part of job performance) in personnel selection research. When evaluating the validity of selection tests and procedures, the accuracy of these validity estimates depends in turn on the accuracy of criterion performance measurement. Accordingly, there is considerable motivation in selection research to obtain reliable and accurate criterion scores for job incumbents participating in the research. Our chapter covers the task performance criterion "space." Chapter 21 in this volume describes citizenship performance criteria. Specific topics covered in this chapter are (a) relatively objective criterion measures, such as work samples, job knowledge tests, and production rates; (b) subjective measures (i.e., ratings of performance), including different rating formats and rater training strategies; (c) dimensionality of job performance; and (d) validity estimates against task performance for several predictor constructs (e.g., ability, personality, etc.).

OBJECTIVE CRITERIA

At first glance, we might assume that objective criterion measures should be preferred over subjective ratings of performance. However, "objective" may be at least in part a misnomer in that judgment often enters into the use of objective criteria. Also, objective measures are notoriously deficient as criteria because they usually tap into only a small part of the total criterion space. Contamination can also be a serious problem with objective measures. For example, factors beyond the control of the job incumbent can influence objective criterion measures. Nonetheless, when they are relevant to important performance requirements and are reasonably reliable and uncontaminated (or when corrections can be made to reduce contamination), objective measures can be useful for measuring some criterion dimensions. In other words, the deficiency issue may not be a problem with objective criteria if we measure well the task performance part of job performance and have other criterion measures evaluating performance in other aspects of the job.

Production Rates

For jobs that have observable, countable products that result from individual performance (e.g., military recruiters or patrol officers who are assigned traffic enforcement duties), a production rate criterion is a compelling bottom-line index of performance. However, as often noted (e.g., Borman, 1991; Guion, 1965), considerable care must be taken in gathering and interpreting production data. For example, work-related dependencies on other employees or on equipment for determining production rates may create bias in these rates. Also, production standards and quota systems (e.g., in call center jobs) can create problems for criterion measurement.

Instability of production rates is another potential problem. Rothe's (1978) extensive research on production workers showed that week-to-week production rates are only moderately reliable. Correlations between successive weeks' production average .75 with incentives, and .53 with no incentives (Rothe, 1978). Longer periods for data collection may be necessary to ensure stable criterion production rates. Most importantly, researchers attempting to derive production criteria should pay special attention to possible contaminating influences, whereby employees have unequal opportunities to produce at the same rate.

Sales

Initially, sales jobs may seem ideally suited for the use of objective criteria as performance measures. Number of sales per unit time, or some similar index of bottom-line sales volume, appear compelling as global, overall performance measures. However, upon closer inspection significant criterion contamination issues are evident for objective sales criteria.

First, summary sales volume measures are a function of individual skill and effort as well as environmental factors beyond the control of the salesperson. In the Campbell, Dunnette, Lawler, and Weick (1970) behavior-performance-effectiveness model, behavior can be characterized as a task statement, a simple description of what an employee might do on the target job, behavior with no evaluative component. Performance is behavior but with an evaluative component, as in a single critical incident (Flanagan, 1954). Finally, effectiveness includes an outcome that is the results of the performance. Thus, in the context of the Campbell et al. model, objective sales volume is an effectiveness measure. Where environmental influences are substantial and unequal in their effect on sales, criterion measurement will be contaminated.

One way to remove contamination is to adjust sales data for factors such as market potential (e.g., Kanfer & Borman, 1987). A practical strategy for making these adjustments is to create norms for stores, sales territories, or for whatever organizational unit provides the appropriate comparison. Then criterion scores for each salesperson can be compared to scores for other salespersons with roughly the same selling-related environment and thus similar opportunities to produce sales.

Unfortunately, an inherent problem with this approach has to do with the norming process. For example, if large sales territories with many salespersons are used to accomplish the norming, there may be meaningful differences within territories with respect to opportunity to perform. If smaller territories are used, then the norms tend to be unstable because the mean sales performance comparison indices are based on too few salespersons. Thus, how one does the adjusting may be as important as whether or not to adjust.

Another "objective" sales criterion that is often used is percent of quota, which presumably controls for environmental factors. Of course, the accuracy of this measure depends on how accurately one defines the environmental factors when setting the quota.

Work Samples

Work sample or performance tests are sometimes developed to provide criteria for selection research. Some argue that work sample tests have the highest fidelity for measuring criterion

performance. In a sense, the argument is compelling: What could be fairer than to assess employees' performance on a job by having them actually perform some of the most important tasks associated with it? Yet evaluation of work samples as criteria is not quite so simple, and their use involves several issues.

One issue in work sample scoring is whether to evaluate products or process relative to work sample performance. In general, tasks associated with products (e.g., troubleshooting a problem with a radio) can be oriented toward either product or process; tasks with no resulting products (e.g., interviewing a job candidate) must be scored according to process considerations. An advantage to scoring products over process is that assessment is typically more objective. However, if the procedures taken to arrive at the product are also important, process assessment is clearly necessary.

Other issues relevant to scoring work samples are germane here. Difficult-to-score process steps are to be avoided. For example, checking and inspecting steps are difficult, if not impossible, to observe. Ill-defined steps and complex steps where an employee can do well on one part of the step but poorly on another should also be avoided.

Still another issue with scoring work samples is the relative merits of pass/fail marks versus performance-level ratings on task steps. Guion (1978) argued for task step performance ratings (e.g., on a 1 = low to 7 = high scale) because they provide more information. Indeed, many steps seem amenable to a continuous performance scale, where such judgments as "more skillful," "faster," and "more efficient" may have meaning for evaluating performance. For certain very simple task steps, pass/fail may suffice, but it will usually be desirable to develop continuous performance scales for use in work sample testing.

A major issue with work sample tests is that researchers may treat them as ultimate criteria; that is, these tests are sometimes considered the criterion of choice for accurately assessing performance in certain jobs, especially those that require complex motor skills. Work samples should not be thought of in this light. First, they are clearly maximum performance rather than typical performance measures. As such, they tap the "can-do" more than the "will-do" performance-over-time aspects of effectiveness. Yet will-do longer-term performance is certainly important for assessing effectiveness in jobs. Accordingly, these measures are deficient when used exclusively in measuring performance. In sum, inherent shortcomings of work samples for measuring some aspects of performance, as well as practical limitations such as time and equipment constraints, argue against relying on such tests to provide a comprehensive index of overall performance.

Job Knowledge Tests

Another category of criterion measures is the job knowledge test. Once the target tasks are identified, items can be prepared, typically in a multiple-choice format, although other kinds of items such as the essay type are of course possible. Just as in writing any other multiple-choice items, care should be taken to ensure that the item stems and response alternatives are clearly stated and that distractor responses are definitely wrong but plausible.

An issue with job knowledge test development is when is the paper-and-pencil knowledge test medium appropriate for evaluating job performance. When a task is procedural, requiring primarily knowledge about steps to complete it, and not complex motor skills for performing each step, a job knowledge format seems clearly to be as appropriate as a work sample format. Tasks requiring certain skills and operations are probably not amenable to job knowledge testing. Such tasks include (a) those that require finely tuned acts of physical coordination (e.g., a police marksmanship task), (b) those that require quick reaction (e.g., typing a letter under time pressure), and (c) those that require complex time-sharing psychomotor performance (e.g., aircraft cockpit simulator tasks).

SUBJECTIVE CRITERIA

Subjective criteria will typically be synonymous with performance ratings. The notion of supervisors or peers providing numerical scores for employees on job-relevant performance areas

is an interesting idea. Ideally, it provides well-informed observers with a means of quantifying their perceptions of individuals' job performance. This is preferable to verbal descriptions of performances because individuals can now be compared in a reasonably straightforward way. The notion can be viewed as analogous to developing structured job analysis questionnaires to take the place of verbal job descriptions for purposes of comparing jobs (McCormick, 1976). In each case, quantification of perceptions clears the way for scientific study of an area that could not be previously studied in this manner.

The emphasis in this section will be on ratings gathered for-research-only as criteria for selection research applications. Although ratings can be generated for purposes of salary administration, promotion decisions, or employee feedback and development, they are not very relevant to personnel selection research. We should add here that recent research and commentaries clearly suggest that performance appraisal systems have very little impact on individual or organizational effectiveness (e.g., Pulakos, 2004; Pulakos & Mueller-Henson, 2015; Pulakos & O'Leary, 2011), further supporting the point that operational performance appraisal ratings should not be used as criteria for test validation research.

For-research-only performance ratings are the most often used criterion measure in I-O psychology. Landy and Farr (1980) refer to several surveys intended to assess how frequently ratings are used as criterion measures in research reports. The percentages reach 75% and higher, suggesting that considerable attention should be paid to this criterion measurement method in the interest of making such measurement as accurate as possible. Issues in using ratings as performance criteria include (a) design of the rating form to be used and (b) type of training to provide to raters.

Rating Formats

Behaviorally Anchored Rating Scales

Smith and Kendall (1963) extended the notion of critical incidents (Flanagan, 1954) by designing a rating format they referred to as behavioral expectation scales, now generally labeled behaviorally anchored rating scales (BARS). Smith and Kendall reasoned that different effectiveness levels on job performance rating scales might be anchored using behavioral examples of incumbent performance. Accordingly, they developed performance rating dimensions, with scaled behavioral examples anchoring the appropriate effectiveness levels on the dimensions.

Essentially, the rater's task is to compare observed job behaviors of the ratee with the behavioral anchors on the scale to assign a rating on that dimension. This was seen as preferable to evaluating a ratee without guidance regarding the effectiveness levels of different scale points. The BARS idea is more than a format; it is a system, or even philosophy (Bernardin & Smith, 1981). For example, ideally raters should record examples of employee work behavior in preparation for assigning performance ratings.

Another positive feature of BARS is that users of the system typically participate in scale development, enhancing the credibility of the format. Further, from a domain sampling perspective, BARS development steps provide an excellent methodology to aid in identifying all important dimensions for a job.

Behavior Summary Scales

In response to a difficulty some raters have had with BARS, that of matching observed ratee performance and the often very specific, low-base-rate behaviors serving as anchors on the scale, Borman (1979) developed the behavior summary scales (BSS) format. With this format, behavioral incidents are first generated targeting a wide range of levels of effectiveness on each dimension, as with BARS. Second, the incidents are retranslated according to dimension membership and level of effectiveness, also as is done with BARS. Finally, the content of all incidents

reliably retranslated into the high-, mid-, and low-effectiveness levels, respectively, is summarized, resulting in the summary scale anchors. These summary anchors represent sometimes four or five effectiveness levels, but the main point is rather than the BARS practice of having individual incidents as scale anchors, BSS has summary anchors capturing the behavioral content of several individual anchors at each level of effectiveness for each dimension.

Regarding raters' use of the BSS, the most important potential advantage is that the behavioral construct underlying each aspect of job performance is made more evident to the rater. Raters do not need to infer the dimensionality from a series of highly specific incidents. The inferential step is accomplished in scale development, in which the behavioral essence from several specific incidents is distilled in each behavior summary statement.

Accordingly, this approach should increase the probability that raters can match observed ratee behavior directly with scaled behavior. That is, by increasing the scope of behavior representing various performance levels on a scale, chances are greater that one of the anchors will accurately describe a ratee's performance on that dimension.

This argument makes good conceptual sense, but in the one format comparison study pitting BARS against a BSS format, there were no consistent differences between these format types with respect to psychometric error or accuracy (Borman, 1979). Thus, the seeming conceptual advantage of BSS may not make much difference in the actual use of the scale.

Behavior Observation Scales

Latham and Wexley (1981) developed the behavior observation scales (BOS) format with favorably worded behavioral statements that the rater responds to by indicating how frequently the ratee exhibits each of these behaviors.

Latham and Wexley (1981) provided a list of advantages of BOS, including (a) BOS are developed from a systematic job analysis; (b) the content of the explicit behavioral items provides an excellent listing of the job's performance requirements in concrete behavioral terms; and (c) item analysis and factor analytic procedures can be more readily applied to BOS ratings than to BARS or BSS data. To these should be added that BOS items appear to cut down on the complexity of inferences necessary to make a rating, although a study by Murphy, Martin, and Garcia (1982) casts some doubt on this point.

Computerized Adaptive Rating Scales

Each of these behavior-based rating format ideas had appealing features. However, the following question arose: Does format make a difference relative to rating errors or the reliability and validity of the ratings generated by raters using the different formats? Not all of the relevant format comparison studies have been conducted, but the studies that have been completed generally show small differences between formats in terms of level of rater errors, reliability, validity, or accuracy. For example, early reviews of format comparison studies (Landy & Farr, 1983; Schwab, Heneman, & DeCotiis, 1975) concluded that the psychometric properties of the BARS format are probably not much better than the psychometric properties of graphic rating scales (GRS, or scales with numerical rather than behavioral anchors). Borman (1979) found only small differences in halo, inter-rater reliability, and validity for BARS, the BSS, and a graphic rating format. Landy and Farr (1980) went so far as to estimate that the variance accounted for in psychometric quality by rating format was as little as 4%. In fact, they called for a "moratorium" on rating format research, citing the largely negative results.

For the next 20 years, Landy and Farr's suggestion was followed for the most part (Farr & Levy, 2007), but it still seems compelling to explore rating format ideas that might result in more reliable and valid judgments about work performance. Small adjustments made to present formats are unlikely to result in higher reliabilities and validities; however, it still seems important to experiment with formats that are fundamentally different from those currently used in

hopes of developing a format more in alignment with raters' cognitive processes or that somehow calibrates raters' perceptions to help them make more precise judgments about observed performance.

One possible idea in the direction of a different rating measurement method started with consideration of Thurstone's (1927) law of comparative judgment in the context of the performance rating process. Thurstone developed a method for scaling stimuli on the basis of paired-comparison judgments. Arguably, his approach places stimuli on an interval scale. In the context of rating job performance, Borman, Buck, Hanson, Motowidlo, Stark, and Drasgow (2001) reasoned that pairs of behavioral statements might be presented to the rater with instructions to pick the statement that is more descriptive of the ratee. If interval-scale judgments of ratee performance levels can be achieved with this method, the paired-comparison judgments may provide ratings that are more precise than those generated by other rating formats that use a linear numerical scale, which arguably provide only ordinal-level measurement. Another idea that might make the paired-comparison format even more effective is to apply an item response theory (IRT) adaptive testing orientation to the method. For example, the rater could be presented with a series of behavioral statement pairs such that responses to each successive pair provide a more precise estimate of ratee performance.

Accordingly, our notion for computerized adaptive rating scales (CARS) was to develop a paired-comparison rating task that used adaptive testing principles to help raters estimate a ratee's performance level through an iterative paired-comparison rating process. The idea was to initially present two behavioral statements associated with a dimension—one reflecting somewhat below average performance and the other reflecting somewhat above average performance. Depending on which statement the rater indicated was more descriptive of the ratee, the rating algorithm, developed subsequently by Stark and Drasgow (2002), selected two additional behavioral statements—one with a scaled effectiveness level somewhat above the effectiveness value of the statement picked first as the more descriptive, and the other with a scaled effectiveness level somewhat below the effectiveness value of that initially chosen statement. The rater's selection of the more descriptive statement for the second paired comparison then revised the initial estimated ratee effectiveness level, and, as before, the algorithm selected two more statements with effectiveness values bracketing the revised estimated performance level. Thus, analogous to adaptive testing, a ratee's "true" effectiveness level was to be estimated in an IRT sense by this iterative paired-comparison rating task that presents in sequence item pairs that maximize the amount of information about performance derived from each choice of an item.

In a laboratory study to evaluate selected psychometric properties of CARS compared to two other formats, videotapes of six office workers were prepared, depicting prescripted levels of performance on three dimensions, and subjects rated these vignettes using the CARS format and one or the other competing formats (graphic or behaviorally anchored rating scales). Results showed 23–37% lower standard errors of measurement for the CARS format. In addition, validity was significantly higher for the CARS format ($d = .18$). Accordingly, in a laboratory study, CARS showed promising results (Borman et al., 2001).

More recently, we developed a CARS system for the Canadian Forces and were able to conduct a field test comparing the standard error of measurement of supervisor and peer ratings on the CARS and behaviorally anchored rating scales (Borman, Kubisiak, & Grossman, 2013). Results showed that the CARS ratings had on average 19.5% lower standard error, compared to BARS ratings, again greater precision for the CARS ratings (i.e., more reliable differentiation between ratees).

One last point about formats: although different format ideas may not make very large differences related to psychometric properties (with the possible exceptions of CARS), well-articulated performance standards for communicating expectations and providing feedback in operational performance management systems can be quite useful. Thus, especially the behavior-based formats can serve this important purpose in organizations. This is important because we should also point out that the more complex formats (i.e., the behavior-based scales) are generally more expensive to develop and thus need more benefits to justify their use.

Rater Training

Rater training provides a promising approach to improving the quality of performance ratings. Two general kinds of training programs have emerged to help raters generate more error-free and accurate ratings (Smith, 1986; Woehr & Huffcutt, 1994). Rater error training seeks simply to alert raters to certain psychometric or perceptual errors such as leniency/severity, halo, restriction-in-range, and similar-to-me effects. Training often takes the form of a brief lecture on or demonstration of each error and training to avoid such errors when making performance ratings.

Frame-of-reference training (Bernardin & Pence, 1980) attempts to convey to raters that performance is multidimensional and to thoroughly familiarize them with the actual content of each performance dimension. Regarding familiarization, examples of different levels of performance on individual dimensions are typically reviewed with raters, along with the "correct" or actual performance levels the examples reflect (e.g., Pulakos, 1984). Practice and feedback for trainees typically rating videotaped performances are important components of this type of training.

Researchers have conducted studies comparing the psychometric properties and accuracy of ratings made by raters trained using one of the approaches just discussed and ratings generated by untrained raters. Results suggest the following conclusions: (a) error training is usually successful in reducing the target psychometric error (Pulakos, 1984); (b) error training does not improve the quality of ratings when inter-rater reliability or accuracy is used as a criterion (e.g., Borman, 1979); and (c) frame-of-reference training increases rating accuracy (Noonan & Sulsky, 2001; Woehr & Huffcutt, 1994).

A useful observation was offered by Bernardin and Pence (1980): Rater error training is successful in reducing the target psychometric response set or error (e.g., halo), but essentially new response sets are imposed on raters (e.g., to eliminate halo, spread out your ratings across dimensions), resulting in no change in accuracy or a reduction in it. Similarly, Borman (1979) suggested that to direct persons to adjust their rating distributions in some manner is relatively easy for training to accomplish; it is much more difficult to train raters to be more accurate. Frame-of-reference training appears to be the best bet to attain this worthwhile goal.

DIMENSIONALITY OF JOB PERFORMANCE

Almost no one doubts that job performance is a multidimensional construct (Campbell, 1990b; Ghiselli, 1956). To identify these multiple categories of performance for a job, I-O psychologists will typically use task analysis, from which clusters of tasks may be derived to define the performance dimensions (McCormick, Jeanneret, & Mecham, 1972), or critical incidents analysis (Flanagan, 1954), which can also result in a set of performance dimensions for a job. With these approaches to identifying performance categories, the dimension sets are likely to be different across target jobs. At one level, this is how it should be. Jobs are often different. However, there is considerable motivation to identify a set of dimensions that represents the performance requirements in common across jobs. Over the past 25 years or so, at least six attempts have been made to develop such dimension sets. In this section of the chapter, we review the methodologies used in these efforts and present the dimension sets. Then, we review similarities and differences among dimension sets and summarize all of the dimension content into a six-category taxonomy.

Campbell, McCloy, Oppler, and Sager (1993)

Campbell, McCloy, Oppler, and Sager (1993) posited eight latent performance categories that summarize the performance requirements of all jobs. The notion was that not every job has as performance requirements all eight of the dimensions, but that for any single job, a subset of these factors (or all eight) are sufficient for describing its performance requirements. Several of these constructs emerged in factor analyses of the performance measures administered in the

Project A research (a large-scale selection and classification study conducted in the U.S. Army; Campbell, 1990a) across the many jobs studied in that program. As examples, for first-tour soldiers in 19 different jobs, technical proficiency, personal discipline, and effort were consistently represented in factor analyses of performance criterion measures. For second-tour noncommissioned officer jobs, a leadership factor was added to the mix of factors emerging. Accordingly, the Campbell et al. (1993) taxonomy has been largely confirmed for several categories, using data that are highly appropriate for testing its generality across a wide variety of supervisory and nonsupervisory jobs.

Thus, the Campbell et al. (1993) dimension system includes these eight dimensions: (1) job-specific technical proficiency, (2) non-job-specific technical proficiency, (3) written and oral communication, (4) demonstrating effort, (5) maintaining personal discipline, (6) facilitating peer and team performance, (7) supervision/leadership, and (8) management/administration. Parts or all of Dimensions 1, 2, and 4–8 were confirmed in Project A, using multiple methods to measure job performance, including hands-on performance tests; job knowledge tests; supervisor and peer ratings on multiple dimensions of performance; administrative measures such as disciplinary cases, awards and commendations, etc.; and a supervisory situational judgment test (SJT) criterion measure. Importantly, all of these criterion measures were developed and then administered to supervisory and nonsupervisory people. The wide variety of criterion measures used to evaluate job performance constructs ensured the criterion space was comprehensively reflected in the coverage of the performance domain. Accordingly, the system seems quite generalizable across different types of jobs and supervisory and nonsupervisory jobs.

Borman and Brush (1993)

In a second approach, focusing on managerial performance, Borman and Brush (1993) inductively derived an 18-dimension taxonomy of performance categories from existing dimension sets taken from empirical studies of managerial performance. In this project, several existing sets of managerial job performance dimensions were gathered from published and unpublished empirical studies, resulting in a total of 187 independent dimensions. These dimensions were then sorted into categories by 25 subject matter experts (SMEs) on the basis of the similarity of the content of each dimension. These 25 sorting solutions were summarized into a single 187-by-187 correlation matrix using a method described by Rosenberg and Sedlak (1972). Finally, the matrix was factor analyzed, resulting in a set of 18 managerial job performance dimensions. These dimensions were further grouped into four "mega-dimensions."

The first mega-dimension, interpersonal skills and communication, consists of those dimensions involving communication skills, maintaining good interpersonal relationships at work, representing the organization to others, and selling/influencing behaviors. Second, the leadership and supervision mega-dimensions includes those dimensions related to guiding, directing, motivating, training, coaching, developing, and coordinating subordinates, as well as providing feedback as needed. Third, technical activities and the "mechanics of management" involve dimensions pertaining to the technical proficiency required for a job, but also those related to managerial tasks such as planning, organizing, decision making, staffing, monitoring, and delegating. Finally, the last mega-dimension, conscientiousness and dependability, consisted of the somewhat more heterogeneous set of dimensions: useful personal behavior and persistence, handling stress, and organizational commitment. This system, at the 18 dimension level, is at a relatively high level of specificity, especially because it covers only management jobs.

Viswesvaran (1993)

Viswesvaran (1993) built upon the lexical hypothesis of Galton (Goldberg, 1993) to develop a taxonomy of general job performance. In this investigation, Viswesvaran compiled and sorted measures of job performance from the literature into summary categories, resulting in

25 conceptually distinct dimensions. Next, correlations between each pair of these 25 dimensions were obtained from studies utilizing these dimensions. These correlations were used in a meta-analysis to determine the true score correlations between the dimensions. Finally, factor analysis was used to analyze these true score correlations and to derive a set of 10 job performance categories.

The dimensions identified in this investigation were intended to summarize overall job performance. Dimensions identified by Viswesvaran (1993) included interpersonal competence, administrative competence, quality, productivity, effort, job knowledge, leadership, compliance/acceptance of authority, communications competence, and an overall job performance dimension.

Borman, Ackerman, and Kubisiak (1994)

Borman, Ackerman, and Kubisiak (1994) incorporated elements of personal construct theory in developing a 12-dimension taxonomy of performance dimensions arguably relevant to all nonmanagement jobs in the U.S. economy. Briefly, personal construct theory posits that, on the basis of their experiences over time, individuals develop categories or dimensions that they use to interpret and make judgments about events or objects, especially other people. Personal construct theorists believe that these categories represent the natural way that people think about their world, again, especially regarding other people (e.g., Adams-Webber, 1979). The Repertory Grid protocol has provided a method for individuals to generate their personal constructs by contrasting different role persons (e.g., mother, best friend). In this application, we were asking supervisor participants to generate their personal constructs related to job performance, what have been referred to as personal work constructs, or "folk theories" of performance (Borman, 1987).

In particular, 81 supervisors representing many different types of jobs and industries (e.g., sales, manufacturing, service sector) generated the names of several effective workers they had worked with and several relatively ineffective workers. The supervisor sample was instructed to select certain pairs of effective and ineffective employees and generate a performance dimension that differentiated the two employees. Sample members prepared a dimension label and a definition of the dimension.

The supervisors generated a total of 176 reasonably nonredundant dimensions and definitions, and similar to the Borman and Brush (1993) research, 12 I-O psychologists sorted these dimensions into categories according to similarity in the performance areas represented, and a 176-by-176 correlation matrix was generated reflecting the relationship between each pair of dimensions. A factor analysis of this matrix revealed a highly interpretable 12-factor solution. The resulting dimension set might be organized hierarchically, similar to Borman and Brush.

First, a grouping of interpersonal and communication dimensions was evident, consisting of the dimensions of communication and cooperation. Next, technical activities related to the job were represented in the dimensions of job knowledge, task proficiency, productivity, and judgment and problem solving. Finally, useful personal behavior and skills included the dimensions of dependability, integrity and professionalism, initiative, adaptability, organization, and safety.

Hunt (1996)

Hunt's intention was to develop a dimension set that reflected important behavioral dimensions of the performance requirements for entry-level jobs. Using a critical incidents approach, Hunt (1996) derived an eight-dimension taxonomy of generic work behaviors focusing on non-job-specific aspects of performance. In this investigation, Hunt used factor analysis to empirically derive dimensions of generic work behaviors from supervisor ratings of employee behaviors. However, contrary to the typical approach in which a single job family or single organization is used, Hunt obtained supervisory ratings for nearly 19,000 employees in 52 different job settings across 36 different companies. Because of the nature of these data (i.e., each data set included a slightly different combination of the behaviors assessed), multiple factor analyses of

the dimension structure could be conducted. First, a sample of data sets was subjected to factor analysis, resulting in several initial dimension structures. Similarities across these initial taxonomies were then cross-validated through the use of additional factor analyses (using hold-out data sets) and SME ratings of agreement.

These analyses resulted in an eight-dimension taxonomy of generic work behavior, including task and citizenship behaviors. Specifically, the dimension structure consisted of two higher-order dimensions and eight second-order dimensions. The higher-order dimension of required performance behaviors included those behaviors required of an employee for continued employment, including the second-order dimensions of attendance, off-task behavior (i.e., effort expended toward non-job-related activities while at work; e.g., goofing off), and employee deviance (a combination of unruliness, theft, and drug misuse). The second higher-order dimension, organizational citizenship behaviors, comprises the second-order dimensions of schedule flexibility and work ethic (a combination of industriousness and thoroughness). Although Hunt identified a ninth specific dimension, adherence to confrontational rules, this factor was posited to be primarily relevant to cash register work, so Hunt omitted it from his model of generic work behavior. In addition to dimensions generated from the supervisory ratings, Hunt also identified four dimensions of generic work behavior through a review of the literature, including teamwork, problem solving, safety, and personal appearance.

Peterson, Mumford, Borman, Jeanneret, and Fleishman (1999)

O*NET's generalized work activities (GWAs; Borman, Jeanneret, Kubisiak, & Hanson, 1996) provide a broad-level overview of job behaviors that are applicable to a wide range of jobs. The GWA framework contains 42 lower-order dimensions that have been summarized into four "highest" order dimensions. *Information Input* describes those GWAs that focus on how and where information is acquired as part of the job, including looking for, receiving, identifying, and evaluating job-related information. *Mental Processes* summarizes those GWAs that involve information and data processing, as well as reasoning and decision making. *Work Output* describes physical activities that get performed on the job, including manual activities and complex and technical activities requiring coordinated movements. Finally, *Interacting with Others* summarizes GWAs that involve interactions with others or supervisory functions, including communicating, interacting, coordinating, developing, managing, advising, and administering.

Between the 4 and 42 levels of GWA dimensions, there is a nine-dimension system that we focus on here. This system is supported by factor analytic studies of the Position Analysis Questionnaire (PAQ) and other job analysis instruments for nonsupervisory jobs (McCormick, 1976), and the Borman and Brush (1993) behavioral dimensions (in turn derived in part from other factor analyses involving management or supervisory jobs). The nine-dimension system includes (1) looking for and receiving job-related information, (2) identifying and evaluating job-related information, (3) information/data processing, (4) reasoning/decision making, (5) performing physical and manual work activities, (6) performing complex/technical activities, (7) communicating/interacting, (8) coordinating/developing/managing/advising others, and (9) administering (see Chapter 40, this volume, for more on O*NET).

Integrating the Job Performance Dimension Taxonomies

Clearly, important similarities exist across the dimensional taxonomies discussed that allow an integration of the dimension systems, but they also point to "outlier" dimensions in some of the taxonomies that are worth noting. Table 20.1 presents a crosswalk of the six dimensional systems, indicating the commonalities and differences across the systems. Then, we provide a summary column, reflecting the common content where it is evident. Also, the rows of Table 20.1 are ordered such that the first row represents the most commonality across systems, the second row has the next most commonality, and so on.

TABLE 20.1

Summary of Six Performance Taxonomies

Campbell, McCloy, Oppler, & Sager (1993)	*Borman & Brush (1993)*	*Viswesvaran (1993)*	*Borman, Ackerman, & Kubisiak (1994)*	*Hunt (1996)*	*Peterson, Mumford, Borman, Jeanneret, & Fleishman (1999)*	*Summary Categories*
• Communication • Facilitating peer and team performance	• Communicating effectively and keeping others informed • Maintaining good working relationships • Selling/influencing • Representing the organization to customers and the public	• Communication competence • Interpersonal competence	• Communication • Cooperation	• Teamwork •	• Communicating and interacting	Communicating and interacting
• Job-specific technical proficiency • Non-job-specific technical proficiency	• Technical proficiency	• Productivity • Job knowledge • Quality • Effort	• Effort and productivity • Job knowledge • Task proficiency	• Thoroughness	• Performing complex and technical activities • Performing physical and manual work activities	Productivity and proficiency
• Demonstrating effort • Maintaining personal discipline	• Persisting to reach goals • Handling crises and stress • Organizational commitment	• Compliance/ acceptance of authority	• Initiative • Adaptability • Safety • Dependability • Integrity and professionalism	• Industriousness • Adherence to confrontational rules • Safety • Schedule flexibility		Conscientiousness and dependability
	• Decision making/ problem solving		• Judgment and problem solving	• Problem solving	• Reasoning and decision making	Problem solving
	• Administration and paperwork • Planning and organizing • Monitoring and controlling resources • Staffing	• Administrative competence	• Organization		• Administering	Organizing and planning

TABLE 20.1 (CONTINUED)

Summary of Six Performance Taxonomies

Campbell, McCloy, Oppler, & Sager	*Borman & Brush*	*Viswesvaran*	*Borman, Ackerman, & Kubisiak*	*Hunt*	*Peterson, Mumford, Borman, Jeanneret, & Fleishman (O*NET)*	*Summary Categories*
• Supervision/leadership • Management/ administration	• Coordinating subordinates and other resources to get the job done • Guiding, directing, and motivating subordinates and providing feedback • Training, coaching, and developing subordinates • Delegating	• Leadership			• Coordinating, developing, managing, and advising	Leadership and supervision
	• Collecting and interpreting data				• Information and data processing • Identifying and evaluating job-relevant information • Looking for and receiving job-related information	Information processing
				• Off-task behavior • Unruliness • Attendance • Drug misuse • Theft		Counterproductive work behaviors

All six dimension sets have content involving communicating and interacting with others, although Hunt (1996) has only a teamwork dimension, so communicating is not explicitly represented in his framework. In Viswesvaran (1993), Borman and Brush (1993), and Borman et al. (1994), communicating and the interpersonal component are represented separately; in Peterson et al. (1999), at the nine-dimension level, the two constructs are combined in a single dimension.

Productivity and proficiency are likewise reflected in all six dimension sets, although the configuration of performance dimension content for this construct is somewhat different across the dimension sets. For example, Viswesvaran (1993) has four of his nine dimensions related to this construct (productivity, job knowledge, quality, and effort), Borman et al. (1994) have 3 of their 12 dimensions (effort and productivity, job knowledge, and task proficiency) in this category, and Peterson et al. (1999) divide the construct into performing complex/technical activities and physical/manual activities.

The third summary construct, conscientiousness and dependability, is more heterogeneous, but five of the six dimension sets are represented in some fashion. The content varies from Hunt's (1996) industriousness and adherence to rules to Borman and Brush's (1993) persisting to reach goals and handling crises, Borman et al.'s (1994) initiative, adaptability, and safety, Campbell et al.'s (1993) personal discipline and effort (this effort dimension is defined less like productivity and more like a personal quality compared with the other two effort dimensions), and Viswesvaran's (1993) compliance dimension.

Problem solving draws on content from four of the six dimension sets. Three of these four (Borman & Brush, 1993; Borman et al., 1994; Peterson et al., 1999) include elements of decision making in addition to problem solving; Hunt's (1996) system defines problem solving more narrowly.

The fifth construct, organizing and planning, also has representation by four of the dimension sets. Because this construct can be seen as in part management-oriented, it is not surprising that Borman and Brush's (1993) managerial taxonomy has several dimensions in this category (i.e., administration and paperwork, planning and organizing, monitoring and controlling resources, and staffing). Viswesvaran (1993), Borman et al. (1994), and Peterson et al. (1999) have a single administering or organizing dimension. Finally, Campbell et al.'s (1993) management/administration dimension is broader than organizing and planning but does contain elements relevant to this construct.

The sixth summary construct is leadership and supervision and is also represented in four of the dimension sets. Again, as might be expected, the Borman and Brush (1993) managerial taxonomy has multiple dimensions in this category (coordinating subordinates; guiding, directing, and motivating subordinates; training, coaching, and developing subordinates; and a delegating dimension). Campbell et al. (1993) have two leadership-related dimensions (supervision/leadership and at least part of management/administration). We should note that the Hunt (1996) and Borman et al. (1994) taxonomies were intended for entry-level and nonmanagement jobs, respectively, and thus would not be expected to contain supervisory or managerial dimensions.

A seventh construct, information processing, had representation from only two systems: information and data processing, identifying and evaluating job-relevant information, and looking for and receiving job-related information from Peterson et al. (1999) and collecting and interpreting data from Borman and Brush (1993). And, Hunt's (1996) dimension set had several dimensions that could be classified as counterproductive work behaviors. These included off-task behavior, unruliness, drug misuse, and theft.

Because these last two categories were relatively idiosyncratic, represented in only one or two of the dimension sets reviewed, we propose that the six summary construct system might be used as a target criterion taxonomy in personnel selection research. Thus, we argue here that in the future, it might be preferable to consider a more multidimensional criterion space rather than just overall job performance in this quest toward systematically studying links between individual predictor and criterion variables in a selection context. All of the six dimension sets reviewed have important strengths. What we are advocating, following Campbell et al. (1993) and Campbell, Gasser, and Oswald (1996), is that the field move toward some performance taxonomy that can be used in personnel selection research to more systematically study empirical links between individual differences and individual performance constructs, as represented by a

job performance taxonomy. A science of personnel selection could benefit greatly from research using a common set of performance constructs to map individual difference (e.g., abilities, personality, vocational interests) and job performance relations (Borman, Hanson, & Hedge, 1997; Campbell, et al., 1996). This approach gets us beyond studying individual differences–overall job performance correlations.

PREDICTORS OF TASK PERFORMANCE DIMENSIONS

A major problem with studying predictor-task performance relationships is that almost all predictor-performance correlation data use overall performance as the criterion data. Fortunately, there is an important exception. Project A, also known as the U.S. Army's Selection and Classification Project, was a large-scale test validation effort conducted by the Army Research Institute (ARI) and three research firms (Campbell, 1990a; Campbell & Zook, 1990; see also Campbell & Knapp; Chapter 40 in this volume). The seven-year effort included data from thousands of participants across a wide range of military occupational specialties (MOS). In addition to its large sample size, Project A measured multiple dimensions of performance, including task performance. Specifically, criteria were carefully developed based on a literature review, the critical incidents technique, and a clear explication of the task domain.

Performance was measured using multiple indices, including hands-on job sample tests, multiple-choice knowledge tests, and supervisor/peer ratings of performance on Behavior Summary Scales. Army-wide and MOS-specific scales were developed, and administrative/archival records were also examined. On the basis of exploratory and confirmatory factor analytic results, five dimensions of performance were specified: (1) core technical proficiency, (2) general soldiering proficiency, (3) effort and leadership, (4) personal discipline, and (5) physical fitness and military bearing. The first two factors—core technical proficiency and general soldiering proficiency—clearly represent task performance constructs; substantial loadings were evident for hands-on performance tests, the job knowledge tests, and supervisor/peer ratings on some of the technical performance dimensions. Thus, we believe these Project A criterion data are ideal for representing a relatively pure measure of task performance. Additionally, data were collected for each participant on the following five major predictor constructs: general cognitive ability, spatial ability, perceptual/psychomotor ability, personality, and vocational interests. Accordingly, the wide array of predictor data in Project A provide relatively comprehensive coverage of predictor-task performance links.

The remainder of the chapter will summarize the correlations obtained in the concurrent validation study part of Project A (Campbell & Zook, 1990). Mean validities are based on data from 4,039 incumbents in nine diverse MOS, including infantryman, cannon crewmember, armor crewman, single channel radio operator, light wheel vehicle mechanic, motor transport operator, administrative specialist, medical specialist, and military police. Validity estimates were corrected for range restriction and criterion unreliability. Despite the exemplary aspects of Project A, one might question the generalizability of the results, given that an Army sample was used. Thus, for each type of predictor, relevant research conducted with other samples is also discussed.

TABLE 20.2

Validities Against Task Performance by Type of Predictor

General Cognitive Ability	*.63–.65*
Spatial Ability	.56–.63
Perceptual/ Psychomotor Ability	.53–.70
Personality	.25
Vocational Interests	.34–.35

Note: Corrected for unreliability of criterion and restriction of range. N = 4,039.

General Cognitive Ability

A large body of literature has examined the link between general cognitive ability and job performance, and findings indicate that it is one of the most robust predictors of performance (Ree & Earles, 1992; Schmidt & Hunter, 1998; see also Chapter 11, this volume). In Project A, general cognitive ability was measured with the Armed Services Vocational Aptitude Battery (ASVAB), in which nine subtests combined to form four composite scores: technical, quantitative, verbal, and speed. Similar to other research in the selection field, the Project A relationships between general cognitive ability and task performance dimensions were strong, with a mean validity of .63 between cognitive ability and core technical proficiency, and .65 between cognitive ability and general soldiering proficiency.

A substantial body of research has also examined the relationship between cognitive ability and overall job performance using nonmilitary samples. Literally thousands of studies have investigated this research question, finding strong correlations between general cognitive ability and job performance across various jobs, companies, and criteria (e.g., Hunter, 1986; Hunter & Schmidt, 1996; Schmidt & Hunter, 1981). Although research conducted on civilian populations report high validity coefficients between job performance and cognitive ability, they are not as high as those reported in Project A.

For example, Hunter and Hunter (1984) summarized the results of 515 validation studies conducted by the U.S. Department of Labor, with more than 32,000 employees in 512 diverse civilian jobs (Hunter, 1980). On the basis of this large-scale meta-analysis, Hunter and Hunter reported validities of .40, .51, and .58 between general cognitive ability and job proficiency for low-, medium-, and high-complexity jobs, respectively. The .51 estimate for medium complexity jobs was recited in Schmidt and Hunter's (1998) seminal article and is frequently referenced in the literature as a point estimate for the relationship between cognitive ability and job performance. More recently, researchers have conducted meta-analyses on the basis of studies conducted in different countries (e.g., the United Kingdom and Germany), reporting relationships of similar magnitude (Berta, Anderson, & Salgado, 2005; Hülsheger, Maier, & Stumpp, 2007). A likely reason for the higher mean validities presented here is that the criterion was task performance rather than overall performance.

Spatial and Perceptual/Psychomotor Ability

In addition to general cognitive ability, Project A examined the relationship between spatial and perceptual/psychomotor and task performance. Spatial ability was measured with the Spatial Test Battery, comprising six paper-and-pencil tests. The six tests—assembling objects, object rotation, mazes, orientation, map, and figural reasoning—were combined to form an overall composite score. Perceptual/psychomotor ability was assessed in a computerized battery of 20 tests, which formed six composite scores: (1) psychomotor, (2) complex perceptual speed, (3) complex perceptual accuracy, (4) number speed and accuracy, (5) simple reaction speed, and (6) simple reaction accuracy. Sample tests include target identification, cannon shoot, and target tracking.

Although lower than with general cognitive ability, the relationships between spatial and perceptual/psychomotor ability and task performance were high. The correlations with core technical proficiency and general soldiering proficiency were .56 and .63 for spatial ability and .53 and .57 for perceptual/psychomotor ability. These mean validities are substantially higher than those reported in other studies, in which overall performance was the criterion. In addition, several meta-analytic studies have examined these relationships, focusing on such specific industries as aviation (Hunter & Burke; 1994; Martinussen, 1996) and craft jobs in the utility field (Levine, Spector, Menon, Narayanan, & Cannon-Bowers, 1996). Across 68 studies on pilot selection, Hunter and Burke reported mean validities of .19 for spatial ability, .32 for gross dexterity; .10 for fine dexterity, and .20 for perceptual speed, correcting for sampling error only. Similarly, Martinussen reported a mean relationship of .20 between psychomotor/information processing and pilot performance. Martinussen's meta-analysis was based on 50 studies conducted in 11

countries. Finally, Levine et al. conducted a meta-analysis of 80 studies that sampled craft jobs in the utility industry across six job families. The weighted average of correlation coefficients was .20 between spatial/psychomotor ability and overall performance. Thus, similar to with cognitive ability, spatial and perceptual/ psychomotor ability correlate considerably higher with task performance than with overall performance.

Personality

Over the past two decades, research on the utility of personality in the selection field has received a great deal of attention (see Chapter 13, this volume). Although early estimates of the relationship between personality and performance were quite low, more recent results have been somewhat more optimistic. Personality researchers generally credit the advent of a well-accepted taxonomy (i.e., the Big Five) and the increased use of validity generalization techniques (e.g., Barrick & Mount, 1991) for the recent positive findings.

Although Project A did not utilize the Five-Factor Model in the measurement of personality, it did find moderate correlations between personality and task performance. Using the Assessment of Background Life Experiences (ABLE), soldiers completed 11 scales (emotional stability, self-esteem, cooperativeness, conscientiousness, nondelinquency, traditional values, work orientation, internal control, energy level, dominance, physical condition). Seven of the scales combined to form four composite scores: adjustment, dependability, achievement orientation, and physical condition. Overall, the mean validities for personality were .25 for both dimensions of task performance (job-specific technical proficiency and non-job-specific technical proficiency). This is similar to the relationship of .27 reported by Barrick, Mount, and Judge (2001) between conscientiousness and overall job performance in their meta-analysis of 15 earlier meta-analyses.

Hurtz and Donovan (2000) also conducted a meta-analysis. They partitioned the criterion domain into three dimensions: task performance, job dedication, and interpersonal facilitation. Their findings indicate the following relationships between the Big Five and task performance: .15 for conscientiousness, .13 for emotional stability, -.01 for openness to experience, .07 for agreeableness, and .06 for extraversion. The multiple R was .19. Finally, a couple of meta-analyses have examined personality-task performance relations at the Big Five facet level. Woo, Chernyshendo, Stark, and Conz (2014) investigated facets of Openness to Experience and found that, of the eight facets studied, six had higher correlations with task performance than did the overall Openness construct. However, the difference was not great (means = .10 versus .07), and neither of these relationships was very high. Similarly, Dudley, Orvis, Lebiecki, and Cortina (2006) examined four facets of Conscientiousness and found somewhat higher relationships between the facets and task performance compared to Global Conscientiousness and performance (means = .22 and .16), but here the difference was largely due to a single facet, Dependability (r = .46).

In sum, personality-task performance relationships are generally low to modest in magnitude.

Vocational Interests

Another set of predictors investigated in Project A was vocational interests. On the basis of Holland's (1966) Basic Interests Constructs, as well as six different organizational climate scales, Project A researchers developed the Army Vocational Interest Career Examination (AVOICE). Twenty-two scales make up AVOICE, forming six composite scores: skilled technical, structural/machines, combat-related, audiovisual arts, food service, and protective services. Across the six composites, vocational interests related to core technical proficiency ($r = .35$) and general soldiering proficiency ($r = .34$). Conversely, Schmidt and Hunter (1998), citing Holland (1986), commented that there is generally no relationship between interests and job performance. Although they considered this a somewhat surprising finding, they hypothesized that interests may affect one's choice of jobs, but once the job is selected, interests do not affect performance.

More recently, Morris (2003) conducted a meta-analysis of 93 studies, reporting a mean corrected correlation of .29 between vocational interests and job performance. Interestingly, larger effect sizes were observed when studies used task performance as the criterion. The reason for this finding is unclear, but it does mirror the Project A results. Specifically, the correlations between vocational interests and performance were higher for task performance dimensions (.34 to .35) compared with the other three performance dimensions (.12 to .24).

Finally, Nye, Su, Rounds, and Drasgow (2012), in a meta-analysis of 60 studies and 568 correlations, found a mean correlation (uncorrected) of .20 between vocational interests and overall job performance.

In sum, Project A research supports quite strong relationships between general cognitive ability, spatial ability, and perceptual/psychomotor ability and task performance. Importantly, these correlations are higher with task performance than when the criterion is overall performance, the criterion almost always used in meta-analyses of these predictors' validities against job performance. The explanation we offered for this finding is the consistent trend in the literature of combining task and other dimensions of performance into one overall factor, thus reducing the validities of these predictors. Relations for personality and vocational interests with task performance are more modest but still far from trivial (mid-.20s for personality and .20 to mid-.30s for vocational interests).

REFERENCES

Adams-Webber, J. (1979). Intersubject agreement concerning relationships between the positive and negative poles of constructs in repertory grid tests. *British Journal of Medical Psychology, 52*, 197–199.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *Personality and Performance, 9*, 9–30.

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.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of Behaviorally Anchored Rating Scales (BARS). *Journal of Applied Psychology, 66*, 458–463.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387–409.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*, 410–421.

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. *Organizational Behavior and Human Decision Processes, 40*, 307–322.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 2, pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C., Ackerman, L. D., & Kubisiak, U. C. (1994). Development of a performance rating program in support of Department of Labor test validation research. Unpublished manuscript.

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*, 1–21.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology, 86*, 965–973.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. In J. T. Spence, J. M. Darley, & D. J. Foss (Eds.), *Annual review of psychology* (Vol. 48, pp. 299–337). Palo Alto, CA: Consulting Psychologists Press.

Borman, W. C., Kubisiak, U. C., & Grossman, M. R. (2013). *Development and field test of the Canadian Forces Computer Adaptive Rating Scales* (Institute Report No. 796). Tampa, FL: Personnel Decisions Research Institutes, Inc.

Borman, W. C., Jeanneret, P. R., Kubisiak, U. C., & Hanson, M. A. (1996). Generalized work activities: Evidence for the reliability and validity of the measures. In N. G. Peterson, M. D. Mumford, W. C. Borman,

P. R. Jeanneret, & E. A. Fleishman (Eds.), *O*NET final technical report*. Salt Lake City: Utah Department of Employment Security.

Campbell, J. P. (1990a). An overview of the Army Selection and Classification Project (Project A). *Personnel Psychology, 43*, 231–239.

Campbell, J. P. (1990b). *The role of theory in industrial and organizational psychology* (Vol. 1, 2nd ed.). Palo Alto, CA: Consulting Psychologist Press.

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

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 255–299). San Francisco: Jossey-Bass.

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., & Zook, L. M. (Eds.) (1990). *Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A* (ARI Research Report 1597). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57.

Farr, J. L., & Levy, P. E. (2007). Performance appraisal. In L. L. Koppes (Ed.), *Historical perspectives in industrial and organizational psychology* (pp. 311–327). Mahwah, NJ: Lawrence Erlbaum.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327–358.

Ghiselli, E. E. (1956). Dimensional problems of criteria. *Journal of Applied Psychology, 40*, 1–4.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26–34.

Guion, R. M. (1965). *Personnel testing*. New York, NY: McGraw-Hill.

Guion, R. M. (1978). *Principles of work sample testing: III: Construction and evaluation of work sample tests*. Alexandria, VA: U.S. Army Research Institute.

Holland, J. (1986). New directions for interest testing. In B. S. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 245–267). Hillsdale, NJ: Lawrence Erlbaum.

Holland, J. L. (1966). *The psychology of vocational choice: A theory of personality types and model environments*. Oxford, England: Blaisdell.

Hülsheger, U. R., Maier, G. W., & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. *International Journal of Selection and Assessment, 15*, 3–18.

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

Hunter, D. R., & Burke, E. F. (1994). Predicting aircraft pilot training success: A meta-analysis of published research. *The International Journal of Aviation Psychology, 4*, 297–313.

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: U.S. Department of Labor, Employment Service.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340–362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law, 2*, 447–472.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–879.

Kanfer, R., & Borman, W. C. (1987). *Predicting salesperson performance: A review of the literature* (Research Note 87–13). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

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

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory and applications*. New York, NY: Academic Press.

Latham, G. P., & Wexley, K. N. (1981). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley.

Levine, E. L., Spector, P. E., Menon, S., Narayanan, L., & Cannon-Bowers, J. (1996). Validity generalization for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human Performance, 9*, 1–22.

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A meta-analysis. *The International Journal of Aviation Psychology, 6*, 1–20.

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 651–696). Chicago: Rand McNally.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology, 56*, 347–368.

Morris, M. A. (2003). A meta-analytic investigation of vocational interest-based job fit, and its relationship to job satisfaction, performance, and turnover. Unpublished doctoral dissertation. University of Houston.

Murphy, K. R., Martin, C., & Garcia, M. (1982). Do behavioral observation scales measure observation? *Journal of Applied Psychology, 67*, 562–567.

Noonan, L., & Sulsky, L. M. (2001). Examination of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance, 14*, 3–26.

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science, 7*(4), 384–403.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than g. *Journal of Applied Psychology, 79*, 845–851.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.) (1999). *The occupation information network (O*NET)*. Washington, DC: American Psychological Association.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*, 581–588.

Pulakos, E. D. (2004). *Performance management: Society for Human Resource Management*. Washington, DC: SHRM Foundation.

Pulakos, E. D., & Mueller-Henson, R. (2015). *A path to performance management reform*. Alexandria, VA: Association for Talent Development Press.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1*, 86–89.

Rosenberg, S., & Sedlak, A. (1972). Structural representations of perceived personality trait relationships. In A. K. Romney, R. N. Shepard, & S. B. Nerlave (Eds.), *Multidimensional scaling* (pp. 134–162). New York, NY: Seminar.

Rothe, H. F. (1978). Output rates among industrial employees. *Journal of Applied Psychology, 63*, 40–46.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36*, 1128–1137.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549–562.

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*, 22–40.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149–155.

Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26*, 208–227.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286.

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

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

Woo, S. E., Chernyshenko, O. S., Stark, S. E., & Conz, G. (2014). Validity of six openness facets in predicting work behaviors: A meta-analysis. *Journal of Personality Assessment, 96*(1), 76–78.