The Modeling and Assessment of Work Performance

John P. Campbell and Brenton M. Wiernik

Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455; email: campb006@umn.edu

Annu. Rev. Organ. Psychol. Organ. Behav. 2015. 2:47–74

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

This article's doi: 10.1146/annurev-orgpsych-032414-111427

Keywords

performance models, contextual performance, organizational citizenship, counterproductive work behavior, performance dynamics, performance appraisal

Abstract

Individual work role performance drives the entire economy. It is organizational psychology and organizational behavior's (OP/OB's) most crucial dependent variable. In this review, alternative specifications for the definition and latent structure of individual performance are reviewed and summarized. Setting aside differences in terminology, the alternatives are remarkably similar. The Campbell (2012) model is offered as a synthesized description of the content of the latent structure. Issues pertaining to performance dynamics are then reviewed, along with the role played by individual adaptability to changing performance requirements. Using the synthesized model of the latent content structure and dynamics of performance as a backdrop, issues pertaining to the assessment of performance are summarized. The alternative goals of performance assessment, general measurement issues, and the construct validity of specific methods (e.g., ratings, simulations) are reviewed and described. Cross-cultural issues and future research needs are noted.

INTRODUCTION

This article addresses the current state of the literature regarding the modeling and assessment of performance in a work role. The focus is on individual performance. It is the basic building block on which the entire economy is based (Kim & Ployhart 2014). Without individual performance there is no team performance, no unit performance, no organizational performance, no economic sector performance, no GDP. Despite its importance, research on performance does not compare in size or scope to research on its antecedents and consequences. Of the 1,914 dependent variables reported in primary empirical research articles in *The Journal of Applied Psychology*, *Personnel Psychology*, and *The Academy of Management Journal* between 2008 and 2014, only 350 (18%) are measures of individual performance at work. Certainly, other dependent variables are extremely important, including individual work satisfaction, commitment, engagement, stress/health, and work/family balance. However, without individual performance, there can be no job to be satisfied with, no organization to be committed to, and no work to balance with family. We should strive to understand individual performance to the fullest extent possible.

We focus on several central themes. In the remainder of this first section, we consider what performance is and what it is not. In the second section, we review the similarity and dissimilarity of alternative content models of performance and argue that the latent structure of performance is invariant across levels, functional specialties, organizations, and cultures. In the third section, we also consider the issues of performance dynamics and adaptability to distinguish between the content of performance and its processes and context. In the fourth section, we consider issues related to the assessment of performance in a work role. We highlight recent advances in the measurement of performance, including the use of ratings, work simulations, and technology-enhanced performance monitoring systems. We describe the potential opportunities and pitfalls various assessment methods offer for providing meaningful performance information for different assessment purposes. In the last section, we consider implications of research findings for practice and areas for future research.

What Performance Is

Until the 1980s, there were virtually no attempts to model individual job performance as a construct. There was only the "criterion problem" (Austin & Villanova 1992), and the objective was to find performance indicators that approximate the "ultimate" criterion as closely as possible. The ultimate criterion was defined as an indicator of an individual's total contribution to the goals of the organization. Unfortunately, no such indicator exists.

The situation began to change during the 1980s. For example, the Army Selection and Classification Project (Project A) was able to systematically select a sample of entry-level technical jobs from a population of jobs, develop over 100 separate indicators of performance for each job, and collect performance data on two cohorts of 10,000 enlisted personnel at three points in time: at the end of training, at the end of their first tour of duty, and near the end of their second tour of duty after they had assumed leadership responsibilities (see Campbell & Knapp 2001). This permitted extensive applications of confirmatory factor analysis to test substantive models of the latent structure of performance. Subsequently, multidimensional models of performance as a construct were discussed by Borman & Motowidlo (1993), Campbell et al. (1993), and Murphy (1989a).

From these sources, a consensus developed that individual job performance should be defined as things that people actually do, actions they take, that contribute to the organization's goals. Someone must identify those actions that are relevant to the organization's goals and those that are

not, regardless of whether they are in a written job description. For those that are relevant, the level of proficiency with which the individual performs them must be scaled. Both the judgment of relevance and the judgment of level of proficiency depend on a specification of the important substantive content-based goals of the organization, not content-free goals such as "making a profit," and there may indeed be multiple goals, goal change, or goal conflict.

Nothing in this definition requires that a set of performance actions be circumscribed by the term job or that they remain static over a significant length of time. Neither does it require that the goals of an organization remained fixed or that a particular management cadre is responsible for determining the organization's goals (a.k.a. "vision"). Neither does it say that actions, or goals, must be described at a certain level of specificity. Consequently, it is not a violation of this definition of performance for individual organization members to decide themselves what actions are most relevant for what they think the organization's goals are, or should be. Individuals can be quite active (Frese 2008) or proactive (Griffin et al. 2007). However, goal choices, and decisions about what actions best serve them, must be legitimized by the stakeholders empowered to do so by the organization's charter. Otherwise, there is no organization. Perhaps the indictment of "conventional" job analysis (see Pearlman & Sanchez 2010) should be that it does not validly reflect current and future goals, and the actions that best serve them, because "job analysts" may not be sufficiently knowledgeable about current and future organizational goals to determine the appropriate performance actions for a particular work role.

What Performance Is Not

The above specification is intended to distinguish clearly between performance itself and (*a*) the determinants of individual differences in performance and (*b*) the outcomes of performance (a.k.a. results, goal achievement, the bottom line). It certainly seems the case that the determinants have received the most research attention in our field. They include such things as individual trait variables (e.g., cognitive abilities, personality, stable motivational dispositions, physical characteristics and abilities), state variables (e.g., relevant knowledge and skill, attitudes, malleable motivational states), and situational characteristics (e.g., the reward structure, managerial and peer leadership), as well as the interactions among them. Campbell et al. (1993) have argued that all of the above must affect performance by influencing three direct determinants operating in real time: role-specific knowledge, skill, and choice behavior regarding the direction, intensity, and duration of effort. The direct determinants totally mediate the effects of everything else. However, knowledge, skill, and choice behavior are not to be confused with performance itself. Performance itself is what directly facilitates achieving the organization's goals. Motowidlo et al. (1997) make a similar argument.

Performance must also be distinguished from the outcomes of performance if the variance in an outcome (e.g., sales, stock price, salary) is due to other factors, in addition to the individual's performance level. As noted below, it may indeed be possible to develop outcome measures that are virtually totally under the control of the individual, in which case the outcome measure is a performance measure. We harp on these issues because the distinctions between performance, its determinants, and its outcomes are often overlooked, particularly with regard to leadership (Campbell 2013b). All three are important.

Performance should also not be confused with such indicators as efficiency or productivity, although they are certainly important. Both imply a ratio of output to input; and while it may be possible to think of the cost of achieving a certain level of performance, as defined above, that is not our concern here. Finally, performance is not synonymous with development, attrition, or promotion, although these things are certainly important as well.

MODELING THE LATENT STRUCTURE OF PERFORMANCE

The recent literature has produced a number of attempts to model the substantive content domains of individual performance. These have been cataloged by Koopmans et al. (2011) and include a variety of models that sometimes use different nomenclatures or represent variables other than performance, as it is defined here. There are many references to productivity, work quality, work quantity, in-role performance, extra-role performance, and a number of synonyms for effort, management, leadership, interpersonal behavior, problem solving, teamwork, adaptability, communication, emotional control, cooperation, development, creativity, etc. The lack of an agreed upon nomenclature is striking, and it gives the impression that attempts to explicate the latent structure of performance are quite helter-skelter. However, we assert that the opposite is true. If this catalog is purged of terms that have no specific content specifications (e.g., problem solving, creativity); that refer to metrics, not behavior (e.g., quantity, quality, productivity); or that refer to prerequisite knowledge, skills, or personality traits, rather than performance itself, there is considerable agreement, despite different terms being used for the same thing. The development of this near consensus went something as follows.

Since the 1980s, a number of investigators have suggested models for the latent structure of performance. Given the population of goal relevant actions or behaviors that an individual could perform in a work role, can they be represented by a meaningful dimension structure that describes the major distinguishable components of performance? The assumption here is that the construct of performance is not unidimensional. Advancing the organization's goals requires different categories of individual actions that can be distinguished on the basis of the content of the behavior that is involved, and it is possible to recover these categories.

Early attempts to identify performance categories were job analysis based (e.g., Brumback & Vincent 1970, Fleishman & Quaintance 1984) and used various methods to cluster job tasks on the basis of their content similarity. However, the task content was focused almost exclusively on what is now termed the technical performance dimension. The major milestones along the way to what we think is a near consensus about the broader structure of individual performance are as follows.

Project A

The first major attempt to identify performance factors based on actual performance assessments was the Project A effort during the late 1980s (see Campbell et al. 2001). That effort produced a five-factor solution for entry-level Army enlisted personnel and a six-factor solution for non-commissioned officers (NCOs). On the basis of cross-validation designs and confirmatory factor analyses, the five-factor and six-factor models were shown to be quite robust across Army occupations and across cohorts separated by three years. Both models contained one factor specific to the Army (physical fitness and military bearing), but the remaining factors were more general, dealing with technical performance, peer leadership, supervisory leadership, extra effort/initiative, and personal discipline. The level of goodness-of-fit was remarkable, even when cross-validating from one cohort to the other. The Project A factor structure was also similar to one specified by Lance and colleagues (1992) using Air Force data.

Campbell et al. (1993) expanded the Project A model to make it more appropriate for non-military jobs. Their model portrayed the latent structure of performance as composed of eight factors: job-specific technical proficiency, non-job-specific technical proficiency, communication, demonstrated effort and initiative, personal discipline, facilitating peer and team performance, supervision/leadership, and management/administration. These dimensions were defined at a level

of specificity that was fairly general, but specific enough to be useful for descriptive purposes. There could be more specific subfactors or higher-order factors.

Core Technical and Contextual Performance

Influenced by Project A and subsequent studies, Borman & Motowidlo (1993, 1997) proposed a model of performance with two general factors, which they labeled core task performance and contextual performance. Core task performance "consists of the activities that transform raw materials into goods and services that are the organization's products," and contextual performance "consists of activities that service and maintain the technical core by replenishing its supply of raw materials, distributing its finished products, or providing important planning, supervising, or staff functions that enable it to function effectively and efficiently" (Motowidlo et al. 1997, p. 75). These definitions were not without ambiguity, much of which was resolved in the Borman & Motowidlo (1997) paper that proposed five subfactors for contextual performance described in behavioral terms. Subsequent research has supported both the distinctiveness of core technical versus contextual performance and the nature of the contextual subfactors (see Conway 1996).

Organizational Citizenship Behavior

Based on management theory, Organ (1988) proposed organizational citizenship behavior (OCB) as a major component of performance and defined it as discretionary behavior, not necessarily part of a job description, that promotes the effective functioning of the organization via being cooperative, helping other people, tolerating less-than-ideal working conditions, going well beyond minimal requirements, identifying with the organization's goals, and participating voluntarily in organizational governance and administration. There is now a relatively large literature on the assessment, prediction, and consequences of OCB that is thoroughly reviewed by Organ et al. (2011). As many as 30 different facets of OCB have been proposed, but six factors seem to reflect their content (Podsakoff et al. 2000). Further, the relevant parties agree that there is virtually complete overlap of the six OCB factors with the subfactors of contextual performance (see Motowidlo 2000, Organ et al. 2011), although there is some mild argument over whose language is the most interesting.

Counterproductive Work Behavior

The Project A performance model included a factor labeled maintaining personal discipline. It involved such things as disciplinary actions, violation of regulations, and assessment of personal misconduct. Its counterpart in civilian life is counterproductive work behavior (CWB), the content of which has been specified by Bennett & Robinson (2000), Berry et al. (2007), Gruys & Sackett (2003), and Dalal (2005). The behavioral definition of CWB is "scalable actions and behaviors that employees engage in that detract from organizational goals or well-being. They include behaviors that bring about undesirable consequences for the organization or its stakeholders" (Ones & Dilchert 2013, p. 645). Two primary subfactors have emerged: deviance directed at the organization (CWB-O; e.g., theft, absence) and deviance directed at other individuals (CWB-I; e.g., bullying, gossiping). Although substance abuse is seemingly directed at the self, factor analytically it clusters with behaviors directed at the organization (Marcus et al. 2002), as the behavioral result of drug use is typically to avoid work or perform poorly or unsafely. Spector and colleagues (2006) went further and decomposed CWB into five specific facets, arguing that the facets have different

antecedents. For each of the two primary subfactors, it might also be possible to specify two subfacets corresponding to approach versus avoidance behaviors. The approach–avoidance distinction is a recurring one in motivation (e.g., Gable et al. 2003) and psychopathology (Markon et al. 2005). For CWB, it is a distinction between actively acting against the organization (property deviance: sabotage, theft, etc.) versus staying away (production deviance: unexcused absence, drug abuse, etc.) and between actively acting against other individuals (personal aggression: bullying, physical abuse, etc.) versus subtly undermining or excluding them (political deviance: gossiping, showing favoritism, etc.). The CWB-I/CWB-O factors have received support in many factor analytic studies, though a strong general CWB factor is present (Ones & Dilchert 2013). Evidence also suggests that CWB are not simply the negative end of positively-stated dimensions such as the OCB/contextual performance factors (Berry et al. 2007, Miles et al. 2002, Ones & Dilchert 2013, Spector et al. 2010), but constitute distinguishable separate dimensions that have different determinants. Measures should avoid including both OCB and CWB items on the same scale (Spector & Cha 2014).

Proficiency, Adaptability, and Proactivity

Griffin and colleagues (2007) presented a model of work role performance that does not focus directly on the latent structure of performance behavior. Instead, it posits a 3×3 classification of work role behaviors in which one dimension represents organizational level (individual, team, organization) and the second dimension progresses from proficiency on prescribed tasks; to proficiency in adapting to changes in individual, team, or organizational requirements; to being proactive in instituting new methods or solutions at the individual, team, or organizational level. Three items (i.e., rating scales) assess proficiency within each of the nine cells. Given the item content and definitions for each cell, the level dimension seems to represent (*a*) individual task performance, (*b*) peer leadership and support in teams, and (*c*) certain aspects of the management role. The proactivity column represents the extra effort/initiative component of several other models. The adaptivity column represents a much more complex construct and will be discussed subsequently. In essence, the Griffin et al. (2007) model represents much of the same content as the models described above, but in a 3×3 format rather than a hierarchical one. Their argument is that it makes the effects of context (i.e., level) clearer.

Competency Models

Competency modeling is an important area of practice in human resource (HR) management (Shippmann 2010), and it has relevance for the specification and assessment of performance, particularly with regard to management performance. Unfortunately, there remains some ambiguity in specifying what a competency is. In our view, there are three choices: A competency could refer to performance itself, to a direct determinant of performance (e.g., negotiating skill), or to a more distal indirect determinant of performance (e.g., openness to experience). Shippmann et al. (2000) seem to allow all three. In an attempt to clarify, Campion et al. (2011) characterize competencies both as KSAOs (determinants of performance) and as "performance capabilities," which also seems to allow all three. Tett et al. (2000) did a content analysis of published competency models, identified 53 competencies, and attempted to define each of the 53, which were grouped into 10 categories. The 53 competencies did not uniformly represent performance itself. Some seemed to represent a necessary skill, and others seemed to represent personality characteristics. Also, the processes by which competencies were named and included in a model were not always very clear.

Stevens (2013) discusses the current state of competency modeling and seems to conclude that a competency model must include both the important determinants of performance and the important factors of performance itself because both are necessary to inform selection, training and development, promotions, job assignment, and compensation. That is, in the best of worlds, all critical HR systems should be aligned with the same competency model. The real issue is how expertly the competencies are specified and assessed.

Bartram (2005) tried to restrict a subset of competencies to performance itself and provided a specification of the "great eight" competencies used by the consulting firm SHL. The great eight competencies are higher-order factors representing 112 individual scales, some of which could be construed to represent knowledge or skill, rather than performance. However, the great eight probably came the closest to making competency synonymous with performance.

The Campbell Revision

Given what has transpired from the late 1980s to the present, Campbell (2012) revised Campbell et al. (1993) to represent a consensus latent structure that is described as concretely as possible. That is, the intent was to use as few difficult-to-define abstractions as possible, even though it makes things sound less exciting.

The eight basic substantive factors of individual performance in a work role are asserted to be the following:

1. **Technical performance:** All models acknowledge that virtually all jobs or work roles have technical performance requirements. Such requirements can vary by substantive area (driving a vehicle versus analyzing data) and by level of complexity or difficulty within area (driving a taxi versus driving a jet liner, tabulating sales frequencies versus modeling institutional investment strategies). As noted by Wisecarver et al. (2007), this factor should also include, what they term, core interpersonal tasks such as those involved when dealing with patients, vendors, customers, or community members. A common term for these tasks is customer service. They are no less technical than maintaining equipment. The subfactors for this dimension are obviously numerous, and the domain could be parsed into wide or narrow slices.
2. **Communication:** The Campbell et al. (1993) model is the only one that isolated communication as a separate dimension, but it appears as a subfactor in virtually all others. It refers to the proficiency with which one conveys information that is clear, understandable, compelling, and well organized. It is defined as being independent of subject matter expertise, and thus a separate factor, and is not limited to formal communication. The two major subfactors are oral and written communication, and their importance can vary widely across work roles.
3. **Initiative, persistence, and effort:** This factor emerged from the contextual performance and management performance literatures, as well as the OCB literature, where it was referred to as conscientious initiative. It was also part of the Project A factor model. To make this factor conform to the definition of performance used in this article, it must be composed of observable actions. Consequently, it is typically specified in terms of working extra hours, voluntarily taking on additional tasks, going beyond prescribed responsibilities, or working under extreme or adverse conditions. Frese (2008) uses the term active performance to describe these kinds of actions.
4. **Counterproductive work behavior:** Consistent with other models, CWB refers to a category of individual actions or behaviors that are under individual control and

have a negative effect on the goals of the unit or organization. As noted above, there seems to be general agreement that CWB has two major subfactors: deviance directed at the organization and deviance directed at other individuals. Also, as explained above, it seems reasonable to expect an approach–avoidance, or moving toward versus moving away, distinction for both organizational deviance and individual deviance.

5. **Supervisory, managerial, executive (i.e., hierarchical) leadership**: This factor refers to leadership performance in a hierarchical relationship. It also distinguishes between leadership and management. Leadership refers to the interpersonal influence process. The substantive content, as specified by the leadership research literature, is most parsimoniously described by six subfactors (see sidebar titled Six Subfactors Comprising Leadership Performance; also Campbell 2012), based on the remarkable convergence of the literature from the Ohio State and Michigan studies through the contingency theories of Fielder, House, Vroom, and Yetton to the current emphasis on being charismatic and transformational, leading the team, and operating in highly complex and dynamic environments. The subfactors describe what leaders do, not the outcomes of performance (e.g., effective leader–member exchange, follower satisfaction, unit profitability) or the determinants (e.g., cognitive ability, personality) of leadership performance or the situational influences on leader performance. The subfactors are not "styles" or ephemeral "perceptions." In a given setting, the relative emphasis across subfactors may be different, and different leadership models may hypothesize different paths from leader performance to leader effectiveness (i.e., outcomes), which for some people may be the interesting part, but the literature's characterization of leader performance itself (including transformational and charismatic leadership) seems always within the boundaries of these six factors (see Campbell 2013b). Similarly, the six subfactors circumscribe hierarchical leadership performance at all levels. However, the relative emphasis on each subfactor may be different at different organizational levels, and the specific actions within each subfactor may also receive differential emphases. It is also the case that individuals may react differentially to interpersonal influence attempts by the "leader." Leadership is a series of reciprocal processes.
6. **Hierarchical management performance**: Within a hierarchical organization, this factor, distinct from leadership as interpersonal influence, includes those actions that deal with generating, preserving, and allocating the organization's resources to best achieve its goals. Given the existing literature, Campbell (2012) argues that there are eight meaningful subfactors (see sidebar titled Eight Subfactors Comprising Management Performance). As is true for the components of leadership, there may be considerably different emphases on the management performance subfactors across work roles and also as a function of the type of organization, organizational level, changes in the situational context, or changes in organizational goals. Also, there can be very high scorers and very low scorers on both the leadership and management subfactors. Very high scorers on certain critical dimensions are transformational (e.g., Walumbwa & Wernsing 2013). Very low scorers are dysfunctional (Hogan et al. 2011).
7. **Peer/team member leadership performance**: The content of this factor is parallel to the actions that comprise hierarchical leadership (Factor 5 above). The defining characteristic is that these actions are in the context of peer or team member interrelationships, and the peer/team relationships in question can be at any organizational level (e.g., production teams versus management teams). Many behaviors that comprise the OCB dimension of personal support (e.g., helping, cooperating, courtesy, motivating) that are not part of hierarchical leadership also belong here.

8. **Peer/team member management performance:** A defining characteristic of the high-performance work team (e.g., Campbell & Kuncel 2001, Goodman et al. 1988) is that team members perform many management functions, such as planning and problem solving, determining within-team coordination requirements and workload balance, and monitoring team performance. In addition, the contextual performance and OCB literatures both strongly indicate that representing the unit or organization to external stakeholders and exhibiting commitment and compliance to the policies and procedures of the organization are critical performance factors at any organizational level. Consequently, to a greater extent than most researchers realize or acknowledge, there are important elements of management performance in the peer or team context as well as in the hierarchical setting.

Again, these eight factors are intended to be an integrative synthesis of what the literature has suggested are the principal content dimensions of performance in a work role. They are meant to encompass all previous work on individual performance modeling, team member performance, and leadership and management performance.

Because the different modeling efforts have had different starting points and relied on different research streams, the degree of similarly across them is remarkable. Consequently, we assert that at a given level of specificity, the eight factors represent a latent structure for performance that is essentially invariant across organizational levels (including the team context), functional specialties, industry sectors, and types of organizations. This invariance does not preclude varying degrees of importance for the factors, or subfactors, as a function of the specific work role, changes in goals, or other properties of the context. However, a clear implication is that selection, training, appraisal, and reward systems should be consistent with this latent structure. It is intended as a universal competency model of performance.

This latent structure does not preclude higher-order factors (e.g., contextual performance/OCB) or more specific subfactors (e.g., the taxonomy of detailed work activities in the O*NET

SIX SUBFACTORS COMPRISING LEADERSHIP PERFORMANCE

1. Consideration, support, person-centeredness: Providing recognition and encouragement, being supportive when people are under stress, giving constructive feedback, helping others with difficult tasks, and building networks with and among others.
2. Initiating structure, guiding, directing: Providing task assignments, explaining work methods, clarifying work roles, and providing tools, critical knowledge, and technical support.
3. Goal emphasis: Encouraging enthusiasm and commitment for the group/organization goals and emphasizing the important missions to be accomplished.
4. Empowerment, facilitation: Delegating authority and responsibilities to others, encouraging participation, and allowing discretion in decision making.
5. Training, coaching: Providing one-on-one coaching and instruction regarding how to accomplish job tasks, how to interact with other people, and how to deal with obstacles and constraints.
6. Serving as a model: Modeling appropriate behavior regarding interacting with others, acting unselfishly, working under adverse conditions, reacting to crisis or stress, working to achieve goals, showing confidence and enthusiasm, and exhibiting principled and ethical behavior.

(Source: Campbell 2012)

EIGHT SUBFACTORS COMPRISING MANAGEMENT PERFORMANCE

1. Decision making, problem solving, and strategic innovation: Making sound and timely decisions about major goals and strategies and forecasting future trends and formulating strategic and innovative goals (a.k.a. vision) to take advantage of them.
2. Goal setting, planning, organizing, and budgeting: Formulating operative goals, determining how to use personnel and resources to accomplish goals, anticipating potential problems, and estimating costs.
3. Coordination: Actively coordinating the work of two or more units or the work of several work groups within a unit; this includes negotiating and cooperating with other units.
4. Monitoring unit effectiveness: Evaluating progress and effectiveness of units against goals, and monitoring costs and resource consumption.
5. External representation: Representing the organization to those not in the organization (e.g., customers, clients, government agencies, nongovernment organizations, the public).
6. Staffing: Procuring and providing for the development of human resources; this does not include one-on-one coaching, training, or guidance.
7. Administration: Performing day-to-day administrative tasks, documenting actions, and making information available in a timely manner.
8. Commitment and compliance: Complying with and showing commitment to the policies, procedures, and directives of the organization, together with providing loyal constructive criticism.

(Source: Campbell 2012)

data model). However, the argument in Campbell (2012) is that aggregating above the eight factor level loses information. For example, peer leadership and peer management are sufficiently distinct that aggregating them into "citizenship performance" may introduce additional ambiguity into the meaning of a particular score. Whether these distinctions can be captured with existing measurement technologies is another matter. Two other issues with which such models must deal are the existence of a general factor and compound factors.

The General Factor

A general factor does, in fact, exist in virtually all performance indicator covariance matrices, especially those generated by performance ratings (Viswesvaran et al. 2005). The general factor can be produced both by a subset of determinants that are common across subfactors (e.g., cognitive ability and conscientiousness, or real-time knowledge and skill) and by systematic measurement errors (e.g., common method variance, common rater bias, implicit performance models). However, the general factor is not a single latent variable and has never been defined as such (Campbell 2013a). It is a formative construct, rather than a reflective construct (Diamantopoulos et al. 2008). That is, the general factor does not reflect a single underlying latent variable that can be specified. It must always be "formed" as a sum-score of different components. For decision purposes (e.g., promote/not promote), an overall score must be obtained by weighting and combining the components (see Borsboom et al. 2003 for further discussion of this issue). Given an empirical general factor, there is still considerable differential performance and differential predictability across subfactors, and the training and development considerations are vastly different.

Compound Performance Domains

The eight factors are somewhat analogous to the Big Five dimensions of personality. In both domains, higher-order factors with less informational content and specific facets with more information are present above and below the designated level of specificity. In the same manner, just as there are compound traits in personality that represent combinations of basic traits, researchers have also proposed compound performance dimensions that contain meaningful sets of work behaviors from several of the eight factors that share a common goal (e.g., promoting environmental sustainability, Ones & Dilchert 2012; or enhancing information technology, Taylor & Todd 1995). These compound performance dimensions cut across several dimensions. For example, environmental performance includes technical behaviors (e.g., installing solar panels), peer support behaviors (e.g., encouraging others to recycle), and counterproductive behaviors (e.g., failing to follow waste disposal procedures). It is meaningful to study these behaviors as a set because they are relevant for a common goal and share common antecedents and outcomes. However, these compound dimensions should not be seen as somehow separate from the proposed latent structure of performance.

CONTENT VERSUS PROCESS IN CONCEPTUALIZING PERFORMANCE

When attempting to summarize efforts to model work performance, it is important to distinguish between specifications for performance content and specifications for the performance process (i.e., the context in which performance takes places and the manner in which performance develops and changes over time). So far, this article has centered on a model of performance content, about which there is virtually a consensus. There is a parallel universe that addresses the parameters of performance dynamics, including such concepts as active, proactive, and adaptive performance. These two areas of inquiry, the latent structure of performance content and the nature of performance dynamics, are not in competition. They address different issues, but each has important implications for the other.

Performance Dynamics

No one seriously argues that individual work performance does not change over time, either because the performance requirements change and individuals respond or because individuals change even when performance requirements do not. Much of this literature is reviewed by Sonnentag & Frese (2012). There are at least three aspects of performance requirements that could change: (*a*) the behavioral and/or cognitive content of the requirements, (*b*) the level of performance expected, and (*c*) the conditions under which a particular level of performance is expected (or some combination of these). If there are interactive effects between individuals and the nature of the work role content changes, then changes in the rank ordering of people over time result from multiple sources. Given the current and future nature of employment, it is reasonable to expect that such things will happen and are happening. Change is complicated.

Much of organizational psychology and organizational behavior (OP/OB) research and practice deals with planned interventions designed to change performance by enhancing the individual knowledge, skill, and motivational determinants of performance, such as training and development, goal setting, feedback, incentives of various kinds, supervision, and so on. Such interventions, with performance requirements held constant, could increase the group mean, have differential effects across people, or both. The aptitude–treatment interaction is always with us. The performance changes produced can be sizable (e.g., Katzell & Guzzo 1983, Locke & Latham

2002). Interventions designed to enhance individual performance determinants can also be implemented by the individual's own processes of self-management and regulation (Lord et al. 2010). That is, as a result of an individual's self-monitoring and self-evaluation of his or her own performance against goals, additional training can be sought (perhaps from coworkers), different performance goals can be self-set, feedback can be sought, and self-efficacy can change. The effectiveness of these self-regulation processes can vary widely across individuals. In addition, if they have the latitude to do so, people can conduct their own job redesign (i.e., change the behavioral content of their work roles) to better utilize their knowledge and skills and increase the effort they are willing to spend. Academics are fond of doing that.

As noted by Sonnentag & Frese (2012), individual performance can also change simply as a function of the passage of time. Of course, time is a surrogate for such things as practice and experience, the aging process, and changes in affective or emotional states (Beal et al. 2005).

It is most likely the case that for any given individual over any given period of time, many of these sources of performance change are operating simultaneously. Performance dynamics are complex, and attempts to model the complexity have taken many forms. For example, there could be characteristic growth curves for occupations (Murphy 1989b), differential growth curves across individuals (Hofmann et al. 1992, Stewart & Nandkeolyar 2007, Zyphur et al. 2008), both linear and nonlinear components for growth curves (Deadrick et al. 1997, Sturman 2003), and cyclical changes resulting from a number of self-regulatory mechanisms (Lord et al. 2010). Empirical demonstrations of each of these have been established.

A very recent, and very thorough, review of theory and research pertaining to within-person dynamics, and their antecedents, is provided by Dalal et al. (2014). The authors outline the implications of within-person variability for both selection and performance assessment. Their message, and the message here, is that performance dynamics are most likely dimension specific. That is, the likelihood and nature of changes in the behavioral content, difficulty level, and situational parameters of performance most likely differ across performance dimensions. Similarly, the determinants of individuals being able to cope with (i.e., adapt to) such dynamics can also differ across performance dimensions (Pulakos et al. 2006). Thinking of performance as one thing is counterproductive.

Performance Adaptability

The concept of adaptability, as a particular kind of performance dynamic, has taken on many meanings in the literature. For a very broad and thorough review of adaptability as a construct, the reader is directed to Baard et al. (2014) and Chan (2014). For the purposes of this article, adaptability refers to being able to deal effectively with some combination of the following: changes in organization goals, changes in individual performance requirements, and changes in the performance environment, which have already been identified or are anticipated. Adaptability can be viewed either as a component of performance itself or as a property of the individual (i.e., a determinant of performance). Ployhart & Bliese (2006) present a discussion of this issue and argue that it is probably more useful to identify the characteristics of the adaptive individual than it is to propose adaptability as a distinct content dimension of performance. One reason is that the general definition of adaptability is not content domain specific, and it has been difficult to provide specifications for adaptability as a performance dimension. The best attempt to do so is by Pulakos et al. (2000), who proposed eight adaptability performance factors, which they regard as part of the latent structure of performance. The factors were obtained by mining a large database of critical incidents of effective and ineffective performance and using systematic SME (subject matter expert) judgments to identify and categorize the incidents that were reflective of adaptation. Consequently, these authors disagree with the Ployhart & Bliese (2006) position.

However, the interpretation of the adaptability factors is not entirely straightforward. Some of them seem to represent general skills (handling stress, solving problems creatively), whereas others could be construed as specific subfactors of performance (handling particular kinds of emergencies). Part of the difficulty is being clear about the distinction between the direct knowledge, skill, and motivational determinants of performance (e.g., knowing how to handle stress), the context in which performance takes place (e.g., stressful situations), and performance itself (e.g., managing negative emotional displays). Research on each is of great value. For example, Ployhart & Bliese (2006) developed a measure, based on the Pulakos et al. (2000) factors, of self-assessed skills, interests, and response tendencies. Pulakos et al. (2002) also developed a self-report measure of prior experiences, skill levels, and interests relative to the eight adaptability performance factors they proposed. The two instruments both assess performance determinants and should show considerable convergent validity.

It would also be useful to first sort the original sample of critical incidents used by Pulakos et al. (2000) into performance dimensions and then identify those that reflect adaptability. Using this framework, the adaptive incidents should reflect exceptionally high performance on each dimension. Again, such a framework regards adaptability as dimension specific and not as separate components of the latent structure of performance.

In sum, work role performance requirements can change over time, sometimes over very short periods of time, as a result of many factors. The message in this article is that the latent structure of individual work performance is multidimensional, and the eight factors discussed above represent a consensus developed over several decades. In terms of investigating such things as (*a*) the nature of adaptive performance, (*b*) the determinants of adaptive performance, (*c*) the correlation of past performance with future performance over time (and the reasons that it increases or decreases), (*d*) characteristic performance growth curves for occupations or individuals, and (*e*) the nature of performance changes across situations and contexts and over time, the research should be dimension specific. Performance is not one thing, and neither is adaptability. Consequently, any procedure for assessing performance must be clear about the behavioral content of what is to be assessed and also about what performance dynamics are to be accounted for.

PERFORMANCE ASSESSMENT

The structure and dynamics of performance are complex, which makes assessment of performance a very difficult enterprise. Much of the difficulty results from the necessity of defining performance as things that people actually do. This specification rules out using existing outcomes (e.g., sales, defects, ROI) as performance criteria if a significant portion of their variance is not controlled by the individual. Now, it is also true that work roles are designed or invented to accomplish organizational goals and influence the bottom line (see Ployhart & Hale 2014), and the causal path is a complicated and interesting one. However, the basic tenet here is that the individual should not be held accountable for outcome determinants over which he or she has no control.

Purposes for Performance Assessment

It is well recognized that performance can be assessed for one or more of several distinct reasons and that the measurement purpose has a substantial influence on the measurement format, the results of the assessment, and the effects of the assessment on subsequent performance (Murphy & Cleveland 1995). Consider just the following potential reasons for performance assessment.

For research purposes. The objective here is to estimate the individual's "true score" on one or more dimensions of performance to evaluate the validity of selection procedures, assess the efficacy of interventions intended to improve performance, or test a variety of theoretically derived hypotheses that feature performance as a key variable. For these purposes, the construct validity of the measure—that is, whether it assesses components of the model described above—is the prime concern. Reliability is also important, but construct validity and reliability are often put at risk because research is costly and not well supported in our field and because thorough performance assessment does not have a high priority in many research studies.

For legal defense of personnel decisions. The objective here is to support the validity of specific personnel decisions for meeting certain legal requirements. Similar to research purposes, the primary concern for these performance assessments is construct validity and measurement reliability, with reliability being particularly important. Given the potential for serious legal and financial consequences of poor validity and reliability documentation, appropriate measurement design is more common in these situations than when assessment is purely for research purposes.

For high-stakes appraisal. The high stakes referred to here are promotion, dismissal, reassignment, and compensation decisions. Both the individual and the organization have vital interests in the results of the assessment and could be expected to pay considerable attention to their respective interests. Considerations of fairness, transparency, accuracy, appraisal goals, and motivation (of both the appraiser and appraisee) become paramount (DeNisi & Pritchard 2006, DeNisi & Sonesh 2011, Murphy & Dechert 2013).

For performance feedback and development. Assuming that high-stakes decisions are not directly involved, the emphasis here is on identifying specific, substantive performance behaviors that need enhancement or improvement. Building on Kluger & DeNisi (1996), the feedback literature stipulates that, for developmental purposes, feedback should be very concrete and very specific to the individual's performance. Assessment of an individual on the general factors or subfactors discussed in the previous section would not be useful, unless accompanied by more within-factor specifics. Also, to the greatest extent possible, the feedback process should avoid overall evaluations of the appraisee or even overall evaluations on particular dimensions, such as technical, communication, or peer leadership performance. Such evaluations risk diverting attention away from specific performance improvements to affective self-evaluations.

For self-managed performance improvement. Here, the performance goals are self-set, but the requirements for feedback on specific actions and avoidance of general evaluations are the same. Individuals must conduct their self-appraisal accordingly.

Each of the above purposes are quite different, but all of them should be guided by the basic requirement that the assessment must consider all of the eight performance factors, at some level of specificity, to be a comprehensive measure of performance.

General Measurement Issues

Any attempt to assess work performance must also consider the following general measurement issues.

Typical versus maximum performance. This distinction refers to the direction, amplitude, and duration of effort focused on task accomplishment. Are they at the levels typically exhibited by

individuals in their work settings, or do they reflect specific conditions that keep attention highly focused, increase effort levels, and maintain higher effort levels for some period of time (DuBois et al. 1993)? Performance under such conditions is designated as maximum performance. If the goal is to assess maximum performance, which may be appropriate for assessing performance capabilities under certain critical conditions (Mangos et al. 2007), then the assessment method must account for the appropriate motivational conditions. However, Dalal et al. (2014) argue that besides greater effort, requirements for maximum performance can also bring additional abilities into play. Assessment must also account for the fact that in any given work role, different performance requirements may have different priorities (Mangos & Arnold 2008) and that requiring greater effort may create unusual pressures that degrade performance on complex tasks (Klehe & Anderson 2007). Also, some performance requirements simply might not be amenable to increases in effort (Sackett 2007), such as those that incorporate the leadership and management factors discussed above. Beus & Whitman (2012) carried out a meta-analysis of the typical/maximum performance literature. The estimated correlation (corrected) between them was .42. Differential prediction of typical versus maximum performance and variables that moderated the relationship between them were also examined. Under certain conditions, the correlation between them can be quite high (e.g., Deadrick & Gardner 2008).

Performance dynamics. As discussed above, performance is not static. It can change because of changes in performance requirements; changes in the individual because of training, goal setting, motivational interventions, affective states, aging, etc.; or changes in situational conditions, such as constraints or opportunities created by coworkers or production practices (Stewart & Nandkeolyar 2007). The assessment method must take these dynamics into account, if necessary, to achieve the measurement purpose. In general, this must be done either by repeated measurements (e.g., Stokes et al. 2010) or by capturing summary judgments of performance change over time. Again, assessment of performance change should be dimension specific. For example, the dynamics of technical performance and team/peer leadership performance are most likely different.

Cross-cultural performance assessment. As described above, Campbell's (2012) eight factors are presented as a general model of the latent structure of job performance that is universal across jobs, organizations, industries, and levels. In addition, we believe that it is universally applicable across cultures. Although the relationships among factors and the relative importance of factors may differ across jobs, we believe that these clusters of work tasks exist to some degree in all jobs around the world. With this point in mind, it is important to understand how specific manifestations of and relationships among performance factors differ across cultures, as well as how assessment practices differ in various contexts. Empirical studies of the structure of performance in non-American, non-European contexts have only recently been undertaken. For example, He (2012), Rotundo & Xie (2013), and Xu et al. (2013) all examined the structure of CWB in Chinese organizations. Although some dimensions manifest differently in China than in Western cultures (e.g., interpersonal aggression was expressed primarily through indirect and political behaviors, rather than through confrontation), in general, the observed structures were remarkably similar to findings from American and European samples. More studies of this nature, especially studies examining the lower-order structure of other dimensions of performance and the relationships among the eight factors in new cultural contexts, are needed.

Research examining cultural differences in performance assessment processes is also relatively new. Festing et al. (2012) provide an overview of comparative studies of performance appraisal practices and considerations that must be made when designing performance evaluation systems across cultural contexts, such as the influence of unique cultural values on evaluation practices. For

example, common Chinese values of modesty, leniency, and *guanxi* (social harmony and consideration) may reduce the objectivity of performance ratings, as both the assessor and assessee use the evaluation process to serve other goals (Barron & Sackett 2008). The most comprehensive examination of international differences in performance appraisal practices was conducted by Peretz & Fried (2012). These authors found substantial variability in the structure and formalization of performance appraisal across the large number of organizations from 21 countries studied. This variability was strongly related to cultural characteristics, such as power distance, collectivism, and future orientation. Additionally, within a country, organizations that had normative performance assessment practices experienced less absenteeism and turnover than did organizations with culturally divergent practices. These results further reinforce the conclusion that performance assessment systems designed for cross-cultural use need to accommodate local practices and preferences to fulfill their purposes.

The distributional properties of performance. Recently, O'Boyle & Aguinis (2012) raised the issue of the distributional properties of individual performance and their implications for performance assessment, prediction, and management. Based on five examples (faculty publication counts, entertainment industry awards, frequency of being elected to state legislatures, and both positive and negative sports performance indicators), they argue that a Pareto distribution fits the data far better than a normal distribution, thus calling a number of statistical estimation methods (e.g., multiple regression) into question. Further, assessment via ratings is inappropriate, in their opinion, because most applications of rating methodologies attempt to force a normal distribution. Their prescriptions are to assess outcomes, not behavior, and to achieve greater differentiation among the highest "elite" performers, if possible. It is the highest performers that facilitate organizational success. These themes are amplified in Aguinis & O'Boyle (2014). Beck et al. (2014) have taken issue with the O'Boyle & Aguinis assertions. They argue that seven conditions must be satisfied before a data distribution can be called a performance distribution. They then analyze several representative data sets, including indicators of sports performance, that meet the requirements, and show that a normal model does fit the data. Their conclusion is that the O'Boyle & Aguinis results are largely the result of statistical artifacts.

Given the distinction we make between performance and its outcomes, perhaps the most salient point here is that a particular outcome distribution could be quite skewed, but the underlying performance distribution can approach normality. For example, only a few golf professionals ever win a major tournament. Most do not (a skewed distribution). However, the distribution of actual scores (performance) is much more symmetrical. Parametric methods should probably not be used with highly skewed outcome distributions. Also, the relationship of performance to the utility of outcomes need not be linear and could take many forms. Finally, individual performance may not be the only determinant of the outcome distribution, and the reliability of the outcome indicator itself is an issue. For example, the outcomes of mutual fund managers appear to have zero year-to-year reliability (Carhart 1997, Sauer 1997).

Performance Assessment Methods

There are a number of methods that attempt to assess individual work performance, and each has its strengths and weaknesses. There is no ultimate criterion or even one best way. In this section, we consider the applicability of different methods for different assessment purposes and the ability of each method to assess performance itself, as well as deal with the measurement issues described above.

Ratings. Performance ratings by supervisors, peers, subordinates, or by oneself are ubiquitous, and this literature has been reviewed extensively (e.g., DeNisi & Sonesh 2011, Levy & Williams 2004, Murphy & Dechert 2013, Woehr & Roch 2012). We highlight only the major issues here.

The overriding issue is whether ratings have construct validity for the purpose(s) for which they are to be used. That is, are individual differences in rating scores reflective of individual differences in performance itself? Fundamental to judgments of construct validity are the specifications for what is being measured. For any given situation, we must specify what performance is, and what it is not, as concretely and completely as possible. This includes considerations of the dynamics to be considered (e.g., being adaptive), the time interval of interest, and the typical versus maximum performance requirements, in addition to the substantive content of the latent structure. Without such specifications, construct valid assessment is problematic.

The kinds of evidence supporting construct validity are: the correspondence between the performance specifications and the instructions given to raters (including the rating format); the operative goals of the rater (i.e., do they correspond to the measurement purpose?); the rater's level of knowledge about what is to be rated; the level of interrater agreement and interrater reliability; the correlations of ratings with performance assessments using alternative methods; the pattern of correlations with other variables; the existence, or nonexistence, of rater biases (halo, leniency, central tendency); and the degree to which ratings are contaminated, or not contaminated, by particular variables (e.g., rater/ratee gender or race, opportunity to observe, the rater's own performance level, rater accountability, the rater's interpersonal relationship with the ratee, rater personality, level of rater effort, the rater's own implicit performance model versus the one prescribed by rating instructions and format, and the impression management strategies of the ratee). All of these have received varying degrees of research attention.

One of the most critical factors is the goal of the rater during high-stakes appraisal (DeNisi & Sonesh 2011, Kozlowski et al. 1998, Levy & Williams 2004, Spence & Keeping 2010). The goal of assessing the ratee's true performance has frequently been shown to be less important than alternative rater goals, such as rewarding or punishing the ratee, conforming to organizational expectations, or advancing the rater's self-interests. Consequently, it could be argued that ratings should not be used as high-stakes appraisals unless the goals and motivation of the raters can be aligned with the goals of accuracy, fairness, and transparency (Pulakos & O'Leary 2011). In general, high-stakes appraisals do not work very well as research criteria, and their construct validity as assessments of performance itself is suspect (Murphy & Dechert 2013).

The use of 360 ratings for feedback and management development purposes is widespread. Most 360 systems are based on competency models that specify the capabilities that a high-functioning manager/executive should have (DeNisi & Kluger 2000). As already noted, competencies can reflect personality characteristics, motivational tendencies, knowledge, skills, dimensions of performance itself, or even outcome measures (e.g., "achieves results"). In practice, they tend to lack substantive specifications, which makes using them for feedback purposes difficult (Campion et al. 2011, Tett et al. 2000), and competency ratings tend to have low interrater agreement. Sanchez & Levine (2009) argue that competency models function better as general goals than as a means for individual assessment.

For ratings as research criteria, perhaps the most contentious argument concerns whether interrater correlations represent the reliability of performance ratings. Murphy (2008) and Murphy & DeShon (2000) argue that they do not, and these authors call the construct validity of ratings into question. Ones et al. (2008) and Schmidt et al. (2000) contend that interrater correlations are the appropriate estimator of performance ratings reliability in most samples used in organizational research. For them, the construct validity of the ratings is a separate issue. Putka & Hoffman (2014) agree with Schmidt and colleagues (2000) for a specific set of conditions but also

elaborate on how characteristics of the measurement design (e.g., fully crossed, nested, partially nested, ill structured), the research questions being asked (e.g., predicting the rank ordering of ratees versus assessing their actual performance level), and the kinds of generalizations desired (e.g., generalizing across raters or generalizing across dimensions) influence the appropriate choice of the reliability estimator. The choice of estimator can influence the value of the reliability estimate. In general, misspecifying the reliability estimator can underestimate reliabilities to varying degrees, but the discrepancies become substantial only in very ill-structured designs. For most (but certainly not all) research designs that have been reported, estimating reliability using the interrater correlation, rather than proposed alternatives that explicitly estimate additional sources of variance, is not likely to yield seriously biased estimates. However, given at least modest reliabilities, the basic question is still whether supervisor and peer ratings reflect individual differences in performance or individual differences in ratees, raters, or their interactions, which are unrelated to performance itself.

There is not space to review the construct validity evidence in detail, but at least the following points are relevant. Ratings do tend to exhibit considerable halo and leniency effects. Whether this is error or valid variance is another matter. However, ratings for research purposes tend to exhibit less rater bias than do ratings for high-stakes purposes. As a cognitive decision-making process, ratings are susceptible to primacy and recency effects and a strong tendency to make judgments quickly using shortcut heuristics (Fisher 2008). Raters must process a lot of information, much of it from memory. Consequently, interrater reliability is higher for nonmanagerial and less complex jobs, where performance is more readily observable and concretely specified (Conway & Huffcutt 1997).

Rater source effects have been studied relatively extensively. In general, self-assessments exhibit greater leniency, less accuracy, and lower correlations with other variables than do other sources (Dunning et al. 2004, Heidemeier & Moser 2009). They seem to have little construct validity. Source effects for supervisors, peers, and subordinates have been reviewed by Hoffman et al. (2010), Hoffman & Woehr (2009), and Lance et al. (2008). The summary conclusions are that although rater source differences are not large, they also do not represent error. Different raters can have different perspectives. Consequently, rater intercorrelations, to some small degree, can be underestimates of rater reliability. However, different sources do not seem to produce different factor structures (Campbell & Knapp 2001, Facteau & Craig 2001). Several researchers have also pointed out the shortcomings of the multitrait-multimethod matrix as a way to model rating variance components when dealing with incomplete designs. Putka & Hoffman (2014) suggest alternatives.

Finally, a great deal of evidence shows that ratings have consistent correlations with other variables (e.g., cognitive ability and personality) and that there are meaningful differential correlations of such variables with ratings of different performance dimensions (e.g., Organ et al. 2011). The research on rater training, particularly frame-of-reference training, shows that such training significantly improves the construct validity of ratings (Noonan & Sulsky 2001, Schleicher et al. 2002), which would not be expected if ratings did not assess performance itself. A recent study by Hoffman et al. (2012) evaluated the use of frame-of-reference scales (FORS), which attempt to provide more complete specifications for the dimensions to be rated, and showed them to have greater construct validity and accuracy than traditional scales. Also, a meta-analysis by Bommer et al. (1995) suggests that although the overall correlations between ratings and alternative "objective" measures of performance are relatively low, when the performance components being assessed are similar, the intercorrelations are higher. This was supported in a comprehensive multimethod study of jet engine mechanic performance (Vance et al. 1988). There is also a modest literature on the relationship of assessment center ratings and performance ratings obtained later (Hermelin et al. 2007). The correlations are reasonably high, even though assessment center

ratings and later performance ratings reflect the maximum versus typical performance distinction (i.e., assessment centers are designed to elicit maximum performance).

In our judgment, the construct validity of performance ratings is relatively substantial, even though the performance construct is poorly specified in many studies and rater biases do exist. Construct validity would be enhanced further to the extent that performance is concretely specified; the specifications are incorporated in rater training and the rating instruments; and raters (*a*) have observed the ratee extensively, (*b*) accept the rating goal and the performance specifications, (*c*) understand the rating instrumentation, (*d*) know they are accountable for rating accuracy, (*e*) have ample time, and (*f*) are sensitive to such contaminants as liking for the ratee. These conditions are probably not met in many data collections.

Samples, simulations, and proxies. The use of work samples and simulations as criterion measures has a long history in applied psychology. For example, performance on work samples constituted a large amount of the data used in Project A to develop the enlisted and NCO performance models. Hunter (1983) discussed the relative construct validity of ratings and work samples as measures of job performance, and Howard (1983) proposed the use of work samples and simulations to evaluate training outcomes. Distinctions are frequently drawn between work samples, in which an individual performs an actual job task using real job materials (e.g., fixing a real engine, processing real client emails), and simulations, in which individuals perform tasks in fabricated situations or with facsimiles of task materials (e.g., driving using a video simulator, role playing a conflict negotiation). Although there are conceptual differences between these forms of assessment, their use as measures of performance is based on the same logic, and, in most cases, the choice of one over the other is one of practicality, rather than conceptual choice. Also in this category of measures are behavioral proxies that attempt to elicit the same performance responses as actual or simulated work tasks, but that do not closely mimic actual job tasks or situations. The primary examples of this kind are assessment center exercises, which are increasingly being used for developmental (Rupp et al. 2006) and even performance evaluation purposes (Riggio et al. 2003). These three methods share most of the same strengths and weaknesses. Consequently, for the purposes of this article, we refer to all of them as simulations.

One of the key advantages of simulations over other measures of job performance is their ability to assess employees' capabilities for performing critical tasks that are otherwise difficult, unethical, or impossible to assess with any frequency. For this purpose, the most sophisticated simulations have been developed for training and evaluating individuals in medical and related professions (Kunkler 2006) and military contexts (e.g., Colegrove & Bennett 2006). These simulations have seen the most use in training situations, where the level of detail facilitates providing feedback on specific behaviors. Simulations are also useful for measuring potential performance in emergency situations (i.e., performing adaptively), such as emergency landings for pilots or crash avoidance for drivers. Although simulations for other forms of performance have been developed (e.g., for management, Halpin & Biggs 2009; teamwork, Heinrichs et al. 2008; communication, O'Neil et al. 1997; leadership performance, Thomas et al. 2001), they are not currently as sophisticated.

Compared with ratings, simulations have the advantage of being potentially more valid assessments of employees' ability to perform at a particular level of proficiency, and they are assumed to be free from the contamination issues of performance ratings. However, these measures can suffer from construct validity issues that are as serious as those faced by ratings. The primary threat is one of construct deficiency. Simulations can be expensive and time consuming to develop and administer. As a result, they typically assess only a few (or one) critical job tasks, typically representing some form of technical performance. To the extent that other factors of performance are important (as they are to some degree for all jobs), simulations lack construct

validity. Simulations can also suffer from criterion contamination if the elicited behaviors do not accurately reflect real performance on the job with fidelity (Lievens & Patterson 2011). For example, if the controls of a driving simulation device respond differently than an actual vehicle, simulator performance is not an accurate measure of actual job performance. Finally, if simulations are scored using observer ratings, they can suffer from the same perceptual and evaluative biases (though not the sampling biases) as other ratings-based measures of job performance.

The issue of typical versus maximum performance is especially salient for the use of simulations as performance measures. Individuals completing a simulation are likely to perform at their maximum capacity because simulations measure performance within a short period of time, leaving little opportunity for the motivational and self-control processes that decrease performance from maximum to typical levels to take place. Even when individuals perform at satisfactory levels in a simulation evaluation, they may not demonstrate this level of performance consistently on the job. As a result, simulations may be more useful for developmental purposes than for high-stakes decision making, unless the behaviors being assessed are always likely to elicit maximum effort (e.g., emergency responses). On the plus side, requirements for adaptive responses can be built into a simulation.

Technology-enhanced assessment. At an increasing rate, popular press and management and business practice publications are drawing attention to the potential uses of advanced technologies and large amounts of diverse, rapidly generated data (so-called "big data") to improve business practices (Lohr 2013, McAfee & Brynjolfsson 2012). Of particular relevance to this article are suggestions that such technologies can enhance or replace other forms of performance assessment (Hunt 2011). Despite the enthusiasm for these new technologies, caution is warranted, as all too often these systems measure outcomes, such as sales volume, rather than performance itself (Cravens et al. 1993). Technology-based performance assessments can be useful, but only if they measure performance that is under individual control. For example, onboard computer systems have been used to track delivery trucks and to assess safe driving behaviors and compliance with delivery protocols (Kargupta et al. 2010, Şimşek et al. 2013). Similarly, electronic recordings of emails and phone conversations can be used to assess call center, customer service, and communication performance, especially when these techniques are combined with audio transcription and text mining software to reduce the need for supervisors to evaluate each communication (Miller 2003). Rapidly delivered data also present new opportunities to provide immediate feedback to employees, such as by presenting employees with a daily or hourly scorecard of a relevant performance metric (e.g., error rate, change in customer numbers) or using wearable technologies that sound alerts when unsafe movements are made. Again, so long as the data provided by these tools are under employee control, they constitute a potentially useful measure of individual performance. However, big data can suffer from the same criterion deficiency and construct validity problems as simulations. On the plus side, with big data it is potentially possible to capture performance dynamics as they occur naturally.

Additional potential problems with using technology-based monitoring systems to provide feedback are concerns about invasions of privacy and feelings of dehumanization (Hunt 2011, Miller 2003). Care must be taken to ensure employee acceptance of the technology as a legitimate source of performance information.

Goal Achievement and Performance Outcomes

Throughout this article, we have emphasized that measures of performance must assess behavior that is under individual control, not more distal performance outcomes. However, in practice,

many organizations simply want to use outcome measures as indicators of performance. Also, describing performance in terms of attaining mutually set or accepted goals can increase goal achievement and the perceived value of the evaluation process (Locke & Latham 2002). As we have stated before, these indicators constitute performance measures only so long as factors outside of the individual's control are substantially removed from consideration. Pulakos & O'Leary (2010) discuss ways in which that can be done. Sales figures may be an appropriate performance indicator when they reflect only differential levels of effort or skill, such as for call center employees requesting donations from telephone numbers assigned at random, or when environmental and task difficulty factors can be controlled for, such as by referencing a particular employee's sales against the norms for economically similar areas. Similar considerations apply for other jobs, including manufacturing, management, and executive jobs. However, there is again the problem of criterion deficiency, and critical parts of nontechnical performance dimensions (e.g., peer leadership and management) may go unassessed.

An unfortunate example of a poorly chosen outcome measure, which has received substantial media and policy attention, is the attempt to use changes in standardized achievement test scores of public school students to assess individual teacher performance. These are the so-called value-added models, which attempt to assess student test score gains as indicators of learning in a specific teacher's classroom and hold the individual teacher responsible, even though the year-to-year assignment of students to teachers is not random and controlling for classroom differences via statistical covariates is highly problematic. These issues have been thoroughly discussed by Haertel (2013), who concludes that the gain scores are saturated with error and irrelevant variance and are not assessments of teacher performance. We agree, as do other measurement professionals and teachers themselves (e.g., Mueller 2011).

In sum, when appropriately chosen, outcome measures are useful primarily for high-stakes decision making. They offer value for feedback and developmental purposes only so far as they provide information on the behavioral changes necessary to improve performance.

CONCLUSIONS AND RECOMMENDATIONS

Between 1980 and today, a near consensus about what performance is has emerged. Performance is not unidimensional and, strictly speaking, should not be used in the singular (e.g., is he/she a high performer?). It is composed of all the individual actions that support or detract from the organization's goals to varying degrees. At a particular level of generality/specificity there is also general agreement about what the major components of job performance content are, although many of us will be reluctant to give up our own labels for them. For example, the romance of OCB is a powerful thing for many. There is also a consensus that individual performance is dynamic, for many different reasons that can be specified, and its dynamic features are most likely different for different performance components. As OP/OB research continues, researchers should situate their work within the well-understood space of the latent structure of performance, rather than attempt to declare that each new construct is wholly distinct from what has come before. Cumulative science demands that future research build upon previously generated knowledge, not disregard it in favor of exciting new terminology.

Despite widespread acceptance of the definition of performance as what the individual actually does, not determinants or outcomes, all too often researchers continue to conflate the three, especially in the areas of teamwork (e.g., DeChurch & Mesmer-Magnus 2010) and leadership (e.g., Lord & Dinh 2014). For meaningful scientific communication to take place, clarity and precision in language are necessary. Performance should be specified in behavioral terms as things that people do. This is not a behaviorist statement.

Performance in a work role is a complex phenomenon, which makes assessment difficult. There is no way to make it simple. Regardless of whether the measurement method consists of ratings, simulations, outcomes under the control of the individual, or big-data capture, the information obtained must correspond to the specifications for what performance is. The consensus model described above is intended to serve as a basic starting point for all performance assessment, and this includes considerations of individual performance trends and responses to changing requirements and goals.

In this regard, each of the named methods has strengths and weaknesses. In our view, the primary needs for future research and development are the following:

1. How can we build on existing rater training methods to better teach raters what performance is, what goals they must have, what they must know about the person being rated, what dynamics (e.g., specific adaptive responses) they should try to account for, what time interval is of interest, and what potential contaminants of their ratings they must manage (not just halo and leniency effects)? Going further, a widely available MOOC (massively open online course) dealing with these issues could potentially benefit many sectors, perhaps even society at large. The need for judgments of one person's performance by others will not go away. Big data will not replace it. We simply must do it better.
2. How, and for what reasons, do raters actually make rating judgments? What information do they use? How do they combine it? What are their operative goals? We need many more protocol analysis studies examining such questions for each of the major rating purposes. This means sitting beside someone, perhaps virtually, and having them talk through what they are doing. It is different than studying ratings as a cognitive process, valuable though that is. Protocol analysis is used extensively in the study of expertise (e.g., Hoffman & Militello 2009), why not here?
3. How can online performance rating forms for research purposes be structured and delivered to avoid careless ratings and to instill the same feelings of value, responsibility, and attention that can be obtained when ratings are completed in the physical presence of a researcher or supervisor? Collecting ratings in person is time consuming and expensive. How can these drawbacks be avoided without substantial loss in data quality?
4. What are the best ways to promote transparency and procedural justice in the performance evaluation process, for both assessors and assessees?
5. How can simulations be used to assess performance on other dimensions of performance, in addition to technical performance? This is happening to some degree, but new technologies make it possible to do much more.
6. How can simulations be used to assess adaptive responses to changing requirements, and not just on the technical dimension?
7. What are the best ways to display and present behavioral performance data to facilitate effective evaluation and decision making? How can research on dashboards and other methods from computer and data science be effectively combined with meaningful performance metrics (Yigitbasioglu & Velcu 2012)?

In sum, the last 100 years have seen a great deal of research and development regarding the determinants of performance, including a wide variety of contextual variables, and OP/OB has made much progress. It is our hope that over the next few years (fewer than 100), more effort will be devoted to explicating and understanding the dependent variable side of the equation—performance itself.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Kylee Bolt, Jeffrey Braun, Marissa Clark, and Shannon Evans for their invaluable assistance in preparing this article.

LITERATURE CITED

Aguinis H, O'Boyle E. 2014. Star performers in twenty-first century organizations. *Pers. Psychol.* 67:313–50

Austin JT, Villanova P. 1992. The criterion problem: 1917–1992. *J. Appl. Psychol.* 77:836–74

Baard SK, Rench TA, Kozlowski SWJ. 2014. Performance adaptation: a theoretical integration and review. *J. Manag.* 40:48–99

Barron LG, Sackett PR. 2008. Asian variability in performance rating modesty and leniency bias. *Hum. Perform.* 21:277–90

Bartram D. 2005. The great eight competencies: a criterion centric approach to construct validation. *J. Appl. Psychol.* 90:1185–203

Beal DJ, Weiss HM, Barros E, MacDermid SM. 2005. An episodic process model of affective influences on performance. *J. Appl. Psychol.* 90:1054–68

Beck JW, Beatty AS, Sackett PR. 2014. On the distribution of job performance: the role of measurement characteristics in observed departures from normality. *Pers. Psychol.* 67:531–66

Bennett RJ, Robinson SL. 2000. Development of a measure of workplace deviance. *J. Appl. Psychol.* 85:349–60

Berry CM, Ones DS, Sackett PR. 2007. Interpersonal deviance, organizational deviance, and their common correlates: a review and meta-analysis. *J. Appl. Psychol.* 92:410–24

Beus JM, Whitman DS. 2012. The relationship between typical and maximum performance: a meta-analytic examination. *Hum. Perform.* 25:355–76

Bommer WH, Johnson JL, Rich GA, Podsakoff PM, MacKenzie SB. 1995. On the interchangeability of objective and subjective measures of employee performance: a meta-analysis. *Pers. Psychol.* 48:587–605

Borman WC, Motowidlo SJ. 1993. Expanding the criterion domain to include elements of contextual performance. See Schmitt & Borman 1993, pp. 71–98

Borman WC, Motowidlo SJ. 1997. Task performance and contextual performance: the meaning for personnel selection research. *Hum. Perform.* 10:99–109

Borsboom D, Mellenberg GJ, van Heerden J. 2003. The theoretical status of latent variables. *Psychol. Rev.* 110:203–19

Brumback GB, Vincent JW. 1970. Factor analysis of work-performed data for a sample of administrative, professional, and scientific positions. *Pers. Psychol.* 23:101–7

Campbell JP. 2012. Behavior, performance, and effectiveness in the twenty-first century. See Kozlowski 2012, pp. 159–96

Campbell JP. 2013a. Assessment in I/O psychology: an overview. See Geisinger et al. 2013, pp. 355–95

Campbell JP. 2013b. Leadership, the old, the new, and the timeless: a commentary. See Rumsey 2013, pp. 401–22

Campbell JP, Hanson MA, Oppler SH. 2001. Modeling performance in a population of jobs. See Campbell & Knapp 2001, pp. 307–34

Campbell JP, Knapp DJ, eds. 2001. *Exploring the Limits in Personnel Selection and Classification*. Mahwah, NJ: Erlbaum

Campbell JP, Kuncel NR. 2001. Individual and team training. In *Handbook of Industrial, Work & Organizational Psychology*, Vol. 1: *Personnel Psychology*, ed. N Anderson, DS Ones, HK Sinangil, C Viswesvaran, pp. 278–313. Thousand Oaks, CA: Sage

Campbell JP, McCloy RA, Oppler SH, Sager CE. 1993. A theory of performance. See Schmitt & Borman 1993, pp. 35–70
Campion MA, Fink AA, Ruggeberg BJ, Carr L, Phillips GM, Odman RB. 2011. Doing competencies well: best practices in competency modeling. *Pers. Psychol.* 64:225–62
Carhart MM. 1997. On persistence in mutual fund performance. *J. Financ.* 52:57–82
Chan D, ed. 2014. *Individual Adaptability to Changes at Work: New Directions in Research*. New York: Routledge
Colegrove CM, Bennett W Jr. 2006. Competency-based training: adapting to warfighter needs. Mesa, AZ: Air Force Res. Lab.
Conway JM. 1996. Additional construct validity evidence for the task/contextual performance distinction. *Hum. Perform.* 9:309–29
Conway JM, Huffcutt AI. 1997. Psychometric properties of multisource performance ratings: a meta-analysis of subordinate, supervisor, peer, and self-ratings. *Hum. Perform.* 10:331–60
Cravens DW, Ingram TN, LaForge RW, Young CE. 1993. Behavior-based and outcome-based salesforce control systems. *J. Mark.* 57:47–59
Dalal RS. 2005. A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *J. Appl. Psychol.* 90:1241–55
Dalal RS, Bhave DP, Fiset J. 2014. Within-person variability in job performance: a theoretical review and research agenda. *J. Manag.* 40(5):1396–436
Deadrick DL, Bennett N, Russell CJ. 1997. Using hierarchical linear modeling to examine dynamic performance criteria over time. *J. Manag.* 23:745–57
Deadrick DL, Gardner DG. 2008. Maximal and typical measures of job performance: an analysis of performance variability over time. *Hum. Resour. Manag. Rev.* 18:133–45
DeChurch LA, Mesmer-Magnus JR. 2010. The cognitive underpinnings of effective teamwork: a meta-analysis. *J. Appl. Psychol.* 95:32–53
DeNisi AS, Kluger AN. 2000. Feedback effectiveness: Can 360-appraisals be improved? *Acad. Manag. Perspect.* 14:129–39
DeNisi AS, Pritchard RD. 2006. Performance appraisal, performance management and improving individual performance: a motivational framework. *Manag. Organ. Rev.* 2:253–77
DeNisi AS, Sonesh S. 2011. The appraisal and management of performance at work. See Zedeck 2011, pp. 255–79
Diamantopoulos A, Riefler P, Roth KP. 2008. Advancing formative measurement models. *J. Bus. Res.* 61(12):1203–18
DuBois CL, Sackett PR, Zedeck S, Fogli L. 1993. Further exploration of typical and maximum performance criteria: definitional issues, prediction, and White-Black differences. *J. Appl. Psychol.* 78:205–11
Dunning D, Heath C, Suls JM. 2004. Flawed self-assessment implications for health, education, and the workplace. *Psychol. Sci. Public Interest* 5:69–106
Facteau JD, Craig SB. 2001. Are performance ratings from different rater sources comparable? *J. Appl. Psychol.* 86:215–27
Farr JL, Tippins NT, eds. 2010. *Handbook of Employee Selection*. New York: Routledge
Festing M, Knappert L, Dowling PJ, Engle AD. 2012. Global performance management in MNEs: conceptualization and profiles of country-specific characteristics in China, Germany, and the United States. *Thunderbird Int. Bus. Rev.* 54:825–43
Fisher CD. 2008. What if we took within-person performance variability seriously? *Ind. Organ. Psychol.* 1:185–89
Fleishman EA, Quaintance MK. 1984. *Taxonomies of Human Performance: The Description of Human Tasks*. Orlando, FL: Academic
Frese M. 2008. The word is out: We need an active performance concept for modern workplaces. *Ind. Organ. Psychol.* 1:67–69
Gable SL, Reis HT, Ward AJ. 2003. Evidence for bivariate systems: an empirical test of appetition and aversion across domains. *J. Res. Personal.* 37:349–72
Geisinger KF, Bracken BA, Carlson JF, Hansen JIC, Kuncel NR, et al., eds. 2013. *APA Handbook of Testing and Assessment in Psychology*, Vol. 1: *Test Theory and Testing and Assessment in Industrial and Organizational Psychology*. Washington, DC: Am. Psychol. Assoc.

Goodman PS, Devadas R, Griffith-Hughson TL. 1988. Groups and productivity: analyzing the effectiveness of self-management teams. In *Productivity in Organizations: New Perspectives from Industrial and Organizational Psychology*, ed. JP Campbell, RJ Campbell, pp. 295–327. San Francisco: Jossey-Bass

Griffin MA, Neal A, Parker SK. 2007. A new model of work role performance: positive behavior in uncertain and interdependent contexts. *Acad. Manag. J.* 50:327–47

Gruys ML, Sackett PR. 2003. Investigating the dimensionality of counterproductive work behavior. *Int. J. Sel. Assess.* 11:30–42

Haertel EH. 2013. *Reliability and validity of inferences about teachers based on student test scores*. William H. Angoff Meml. Lect. Ser. 14, Cent. Res. Hum. Cap. Educ., ETS Res. Dev., Mar. 22, Washington, DC

Halpin AL, Biggs WD. 2009. Evaluating business plans in a simulation environment. *Dev. Bus. Simul. Exp. Learn.* 36:149–54

He P. 2012. Counterproductive work behavior among Chinese knowledge workers. *Int. J. Sel. Assess.* 20:119–38

Heidemeier H, Moser K. 2009. Self–other agreement in job performance ratings: a meta-analytic test of a process model. *J. Appl. Psychol.* 94:353–70

Heinrichs WL, Youngblood P, Harter PM, Dev P. 2008. Simulation for team training and assessment: case studies of online training with virtual worlds. *World J. Surg.* 32:161–70

Hermelin E, Lievens F, Robertson IT. 2007. The validity of assessment centres for the prediction of supervisory performance ratings: a meta-analysis. *Int. J. Sel. Assess.* 15:405–11

Hoffman BJ, Gorman CA, Blair CA, Meriac JP, Overstreet B, Atchley EK. 2012. Evidence for the effectiveness of an alternative multisource performance rating methodology. *Pers. Psychol.* 65:531–63

Hoffman BJ, Lance CE, Bynum B, Gentry WA. 2010. Rater source effects are alive and well after all. *Pers. Psychol.* 63:119–51

Hoffman BJ, Woehr DJ. 2009. Disentangling the meaning of multisource performance rating source and dimension factors. *Pers. Psychol.* 62:735–65

Hoffman RR, Militello LG. 2009. *Perspectives on Cognitive Task Analysis*. New York: Psychol. Press, Taylor & Francis

Hofmann DA, Jacobs R, Gerras SJ. 1992. Mapping individual performance over time. *J. Appl. Psychol.* 77:185–95

Hogan J, Hogan R, Kaiser RB. 2011. Management derailment. See Zedeck 2011, pp. 555–76

Howard A. 1983. Work samples and simulations in competency evaluation. *Prof. Psychol. Res. Pract.* 14:780–96

Hunt ST. 2011. Technology is transforming the nature of performance management. *Ind. Organ. Psychol.* 4:188–89

Hunter JE. 1983. A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In *Performance Measurement and Theory*, ed. FJ Landy, S Zedeck, JN Cleveland, pp. 257–66. Mahwah, NJ: Erlbaum

Kargupta H, Sarkar K, Gilligan M. 2010. MineFleet: an overview of a widely adopted distributed vehicle performance data mining system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 16th, Washington, DC, July 25–29*, pp. 37–46. New York: ACM

Katzell RA, Guzzo RA. 1983. Psychological approaches to productivity improvement. *Am. Psychol.* 38:468–72

Kim Y, Ployhart RE. 2014. The effects of staffing and training on firm productivity and profit growth before, during, and after the Great Recession. *J. Appl. Psychol.* 99:361–89

Klehe U-C, Anderson N. 2007. Working hard and working smart: motivation and ability during typical and maximum performance. *J. Appl. Psychol.* 92:978–92

Kluger AN, DeNisi A. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol. Bull.* 119:254–84

Koopmans L, Bernaards CM, Hildebrandt VH, Schaufeli WB, de Vet Henrica CW, van der Beek AJ. 2011. Conceptual frameworks of individual work performance: a systematic review. *J. Occup. Environ. Med.* 53:856–66

Kozlowski SWJ, ed. 2012. *The Oxford Handbook of Organizational Psychology*, Vol. 1. New York: Oxford Univ. Press

Kozlowski SWJ, Chao GT, Morrison RF. 1998. Games raters play: politics, strategies, and impression management in performance appraisal. In *Performance Appraisal: State of the Art in Practice*, ed. JW Smither, pp. 163–205. San Francisco: Jossey-Bass

Kunkler K. 2006. The role of medical simulation: an overview. *Int. J. Med. Robotics Comput. Assist. Surg.* 2:203–10
Lance CE, Hoffman BJ, Gentry WA, Baranik LE. 2008. Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Hum. Resour. Manag. Rev.* 18:223–32
Lance CE, Teachout MS, Donnelly TM. 1992. Specification of the criterion construct space: an application of hierarchical confirmatory factor analysis. *J. Appl. Psychol.* 77:437–52
Levy PE, Williams JR. 2004. The social context of performance appraisal: a review and framework for the future. *J. Manag.* 30:881–905
Lievens F, Patterson F. 2011. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *J. Appl. Psychol.* 96:927–40
Locke EA, Latham GP. 2002. Building a practically useful theory of goal setting and task motivation: a 35-year odyssey. *Am. Psychol.* 57:705–17
Lohr S. 2013. Big data, trying to build better workers. *New York Times*, Apr. 21, p. BU4
Lord RG, Diefendorff JM, Schmidt AM, Hall RJ. 2010. Self-regulation at work. *Annu. Rev. Psychol.* 61:543–68
Lord RG, Dinh JE. 2014. What have we learned that is critical in understanding leadership perceptions and leader-performance relations? *Ind. Organ. Psychol.* 7:158–77
Mangos PM, Arnold RD. 2008. Enhancing military training through the application of maximum and typical performance measurement principles. *Perform. Improv.* 47:29–35
Mangos PM, Steele-Johnson D, LaHuis D, White ED. 2007. A multiple-task measurement framework for assessing maximum-typical performance. *Hum. Perform.* 20:241–58
Marcus B, Schuler H, Quell P, Hümpfner G. 2002. Measuring counterproductivity: development and initial validation of a German self-report questionnaire. *Int. J. Sel. Assess.* 10:18–35
Markon KE, Krueger RF, Watson D. 2005. Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. *J. Personal. Soc. Psychol.* 88:139–57
McAfee A, Brynjolfsson E. 2012. Big data: the management revolution. *Harv. Bus. Rev.* 90:60–68
Miles DE, Borman WE, Spector PE, Fox S. 2002. Building an integrative model of extra role work behaviors: a comparison of counterproductive work behavior with organizational citizenship behavior. *Int. J. Sel. Assess.* 10(1/2):51–57
Miller JS. 2003. High tech and high performance: managing appraisal in the information age. *J. Labor Res.* 24:409–24
Motowidlo SJ. 2000. Some basic issues related to contextual performance and organizational citizenship behavior in human resource management. *Hum. Resour. Manag. Rev.* 10:115–26
Motowidlo SJ, Borman WC, Schmit MJ. 1997. A theory of individual differences in task and contextual performance. *Hum. Perform.* 10:71–83
Mueller L. 2011. How I-O can contribute to the teacher evaluation debate: a response to Lefkowitz. *TIP* 49:17
Murphy KR. 1989a. Dimensions of job performance. In *Testing: Theoretical and Applied Perspectives*, ed. RF Dillon, JW Pellegrino, pp. 218–47. New York: Praeger
Murphy KR. 1989b. Is the relationship between cognitive ability and job performance stable over time? *Hum. Perform.* 2:183–200
Murphy KR. 2008. Explaining the weak relationship between job performance and ratings of job performance. *Ind. Organ. Psychol.* 1:148–60
Murphy KR, Cleveland JN. 1995. *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*. Thousand Oaks, CA: Sage
Murphy KR, Dechert PJ. 2013. 2013 performance appraisal. See Geisinger et al. 2013, pp. 611–27
Murphy KR, DeShon R. 2000. Interrater correlations do not estimate the reliability of job performance ratings. *Pers. Psychol.* 53:873–900
Noonan LE, Sulsky LM. 2001. Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Hum. Perform.* 14:3–26
O'Boyle E, Aguinis H. 2012. The best and the rest: revisiting the norm of normality of individual performance. *Pers. Psychol.* 65:79–119
O'Neil HF, Allred K, Dennis RA. 1997. Validation of computer simulation for assessment of interpersonal skills. In *Workforce Readiness: Competencies and Assessments*, ed. HF O'Neil, pp. 229–54. Mahwah, NJ: Erlbaum

Ones DS, Dilchert S. 2012. Employee green behaviors. In *Managing Human Resources for Environmental Sustainability*, ed. SE Jackson, DS Ones, S Dilchert, pp. 85–116. San Francisco: Jossey-Bass/Wiley
Ones DS, Dilchert S. 2013. Counterproductive work behaviors: concepts, measurement, and nomological network. See Geisinger et al. 2013, pp. 643–59
Ones DS, Viswesvaran C, Schmidt FL. 2008. No new terrain: reliability and construct validity of job performance ratings. *Ind. Organ. Psychol.* 1:174–79
Organ DW. 1988. *Organizational Citizenship Behavior: The Good Soldier Syndrome*. Lexington, MA: Lexington Books
Organ DW, Podsakoff PM, Podsakoff NP. 2011. Expanding the criterion domain to include organizational citizenship behavior: implications for employee selection. See Zedeck 2011, pp. 281–323
Pearlman K, Sanchez JI. 2010. Work analysis. See Farr & Tippins 2010, pp. 73–98
Peretz H, Fried Y. 2012. National cultures, performance appraisal practices, and organizational absenteeism and turnover: a study across 21 countries. *J. Appl. Psychol.* 97:448–59
Ployhart RE, Bliese PD. 2006. Individual adaptability (I-ADAPT) theory: conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. See Salas 2006, pp. 3–39
Ployhart RE, Hale D. 2014. The fascinating psychological microfoundations of strategy and competitive advantage. *Annu. Rev. Organ. Psychol. Organ. Behav.* 1:145–72
Podsakoff PM, MacKenzie SB, Paine JB, Bachrach DG. 2000. Organizational citizenship behaviors: a critical review of the theoretical and empirical literature and suggestions for future research. *J. Manag.* 26:513–63
Pulakos ED, Arad S, Donovan MA, Plamondon KE. 2000. Adaptability in the workplace: development of a taxonomy of adaptive performance. *J. Appl. Psychol.* 85:612–24
Pulakos ED, Dorsey DW, White SS. 2006. Adaptability in the workplace: selecting an adaptive workforce. See Salas 2006, pp. 41–71
Pulakos ED, O'Leary RS. 2010. Defining and measuring results of workplace behavior. See Farr & Tippins 2010, pp. 513–29
Pulakos ED, O'Leary RS. 2011. Why is performance management broken? *Ind. Organ. Psychol.* 4:146–64
Pulakos ED, Schmitt N, Dorsey DW, Arad S, Borman WC, Hedge JW. 2002. Predicting adaptive performance: further tests of a model of adaptability. *Hum. Perform.* 15:299–323
Putka DJ, Hoffman BJ. 2014. "The" reliability of job performance ratings equals 0.52. In *More Statistical and Methodological Myths and Urban Legends*, ed. CE Lance, RJ Vandenberg, pp. 247–75. New York: Taylor & Francis
Riggio RE, Mayes BT, Schleicher DJ. 2003. Using assessment center methods for measuring undergraduate business student outcomes. *J. Manag. Inq.* 12:68–78
Rotundo M, Xie JL. 2013. Understanding the domain of counterproductive work behaviour in China. In *Human Resource Management "With Chinese Characteristics": Facing the Challenges of Globalization*, ed. M Warner, pp. 86–107. New York: Routledge
Rumsey MG, ed. 2013. *The Oxford Handbook of Leadership*. New York: Oxford Univ. Press
Rupp DE, Gibbons AM, Baldwin AM, Snyder LA, Spain SM, et al. 2006. An initial validation of developmental assessment centers as accurate assessments and effective training interventions. *Psychol. Manag. J.* 9:171–200
Sackett PR. 2007. Revisiting the origins of the typical-maximum performance distinction. *Hum. Perform.* 20:179–85
Salas E, ed. 2006. *Advances in Human Performance and Cognitive Engineering Research*, Vol. 6. Bingley, UK: Emerald
Sanchez JI, Levine EL. 2009. What is (or should be) the difference between competency modeling and traditional job analysis? *Hum. Resour. Manag. Rev.* 19:53–63
Sauer DA. 1997. Information content of prior period mutual fund performance rankings. *J. Econ. Bus.* 49:549–67
Schleicher DJ, Day DV, Mayes BT, Riggio RE. 2002. A new frame for frame-of-reference training: enhancing the construct validity of assessment centers. *J. Appl. Psychol.* 87:735–46
Schmidt FL, Viswesvaran C, Ones DS. 2000. Reliability is not validity and validity is not reliability. *Pers. Psychol.* 53:901–12

Schmitt N, Borman WC, eds. 1993. *Personnel Selection in Organizations*. San Francisco: Jossey-Bass
Shippmann JS. 2010. Competencies, job analysis, and the next generation of modeling. In *Handbook of Workplace Assessment*, ed. JC Scott, DH Reynolds, pp. 197–232. San Francisco: Jossey-Bass/Wiley
Shippmann JS, Ash RA, Batjtsta M, Carr L, Eyde LD, et al. 2000. The practice of competency modeling. *Pers. Psychol.* 53:703–40
Şimşek B, Pakdil F, Dengiz B, Testik MC. 2013. Driver performance appraisal using GPS terminal measurements: a conceptual framework. *Transp. Res. Part C Emerg. Technol.* 26:49–60
Sonnentag S, Frese M. 2012. Dynamic performance. See Kozlowski 2012, pp. 548–78
Spector PE, Bauer JA, Fox S. 2010. Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *J. Appl. Psychol.* 95:781–90
Spector PE, Cha XX. 2014. Re-examining citizenship: how the control of measurement artifacts affects observed relationships of organizational citizenship behavior and organizational variables. *Hum. Perform.* 27:165–82
Spector PE, Fox S, Penney LM, Bruursema K, Goh A, Kessler S. 2006. The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? *J. Vocat. Behav.* 68:446–60
Spence JR, Keeping LM. 2010. The impact of non-performance information on ratings of job performance: a policy-capturing approach. *J. Organ. Behav.* 31:587–608
Stevens GW. 2013. A critical review of the science and practice of competency modeling. *Hum. Resour. Dev. Rev.* 12:86–107
Stewart GL, Nandkeolyar AK. 2007. Exploring how constraints created by other people influence intraindividual variation in objective performance measures. *J. Appl. Psychol.* 92:1149–58
Stokes CK, Schneider TR, Lyons JB. 2010. Adaptive performance: a criterion problem. *Team Perform. Manag.* 16(3/4):212–30
Sturman MC. 2003. Searching for the inverted U-shaped relationship between time and performance: meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *J. Manag.* 29:609–40
Taylor S, Todd PA. 1995. Understanding information technology usage: a test of competing models. *Inf. Syst. Res.* 6:144–76
Tett RP, Guterman HA, Bleier A, Murphy PJ. 2000. Development and content validation of a "hyperdimensional" taxonomy of managerial competence. *Hum. Perform.* 13:205–51
Thomas JL, Dickson MW, Bliese PD. 2001. Values predicting leader performance in the U.S. Army Reserve Officer Training Corps Assessment Center: evidence for a personality-mediated model. *Leadersh. Q.* 12:181–96
Vance RJ, MacCallum RC, Coovert MD, Hedge JW. 1988. Construct validity of multiple job performance measures using confirmatory factor analysis. *J. Appl. Psychol.* 73:74–80
Viswesvaran C, Schmidt FL, Ones DS. 2005. Is there a general factor in job performance ratings? A meta-analytic framework for disentangling substantive and error influences. *J. Appl. Psychol.* 90:108–31
Walumbwa FO, Wernsing T. 2013. From transactional and transformational leadership to authentic leadership. See Rumsey 2013, pp. 392–400
Wisecarver MM, Carpenter TD, Kilcullen RN. 2007. Capturing interpersonal performance in a latent performance model. *Mil. Psychol.* 19:83–101
Woehr DJ, Roch S. 2012. Supervisory performance ratings. In *The Oxford Handbook of Personnel Selection and Assessment*, ed. N Schmitt, pp. 517–31. New York: Oxford Univ. Press
Xu S, Wang Q, Liu C, Li Y, Ouyang K. 2013. Content and construct of counterproductive work behavior in a Chinese context. *Soc. Behav. Personal. Int. J.* 41:921–32
Yigitbasioglu OM, Velcu O. 2012. A review of dashboards in performance management: implications for design and research. *Int. J. Account. Inf. Syst.* 13:41–59
Zedeck S, ed. 2011. *APA Handbook of Industrial and Organizational Psychology*, Vol. 2: *Selecting and Developing Members for the Organization*. Washington, DC: Am. Psychol. Assoc.
Zyphur MJ, Chaturvedi S, Arvey RD. 2008. Job performance over time is a function of latent trajectories and previous performance. *J. Appl. Psychol.* 93:217–24

Annual Review of Organizational Psychology and Organizational Behavior

Volume 2, 2015

Contents

Errata

An online log of corrections to *Annual Review of Organizational Psychology and Organizational Behavior* articles may be found at http://www.annualreviews.org/errata/orgpsych.