22

Work Simulations

DEBORAH L. WHETZEL
Human Resources Research Organization

MICHAEL A. MCDANIEL
Virginia Commonwealth University

JEFFREY M. POLLACK
University of Richmond

Imagine that you are applying for a job as an administrative assistant. The employer might interview you or administer tests of job knowledge or personality. The employer might also have you participate in a simulated work day. You are escorted to a desk with a phone and a computer and are told the simulation has begun. Soon, holographic images of your supervisor and coworkers appear, much like a science-fiction movie, and you begin your simulated day interacting with your computer-generated colleagues. The simulation is being scored to determine how well you handle the job requirements. Good luck; your evaluation has started.

Increasingly, employers are turning to work simulations for both selection and training purposes. Although advanced technology is not needed to create and administer simulations for most jobs, technological advances are making sophisticated work simulations more common and economical to develop. Multiple options already exist to select and train employees using advanced technologies (ICT Results, 2008). Employers currently use various in-house options or commercial vendors that offer such services (Employment Technologies Corporation, 2008; Furst Person, 2004; Houran, 2007).[1]

In this chapter, we discuss (a) what work simulations are, (b) the role of work analysis in work simulations, (c) their advantages and limitations, (d) their psychometric properties (e.g., validity, reliability, and subgroup differences), (e) how to create simulations (e.g., development and scoring methods), and (f) future potential developments. Included in our discussion are common concerns, references to useful resources, and integrative examples indicative of main principles.

WHAT ARE WORK SIMULATIONS?

Work simulations are methods of evaluating examinees' performance on tasks that are psychologically or physically similar to what they would do on the job (Callinan & Robertson, 2000; Ployhart, Schneider, & Schmitt, 2006). Depending on one's perspective, this definition applies to a wide

[1] The authors have no relationship with these vendors and neither lobby for nor against using the companies and/or products listed. The authors only use the companies and products as examples.

range of procedures including assessment centers, work sample tests, performance tests, competency tests, and situational judgment tests (McDaniel, Hartman, Whetzel, & Grubb, 2007; Truxillo, Donahue, & Kuang, 2004), and it is consistent with definitions used in the literature (Bobko, Roth, & Buster, 2005; Felker, Curtin, & Rose, 2007; Roth, Bobko, & McFarland, 2005; Truxillo et al., 2004).

Work simulations can be categorized by various characteristics, as described by Callinan and Robertson (2000). Fidelity refers to the extent to which the simulation is similar to the job. Some authors have distinguished between psychological fidelity and physical fidelity (Binning & Barrett, 1989; Goldstein, Zedeck, & Schneider, 1993). Psychological fidelity exists to the extent that the test samples the job-related knowledge, skills, and abilities required for essential job duties. Physical fidelity is the extent to which the test simulates the actual job tasks. Callinan and Robertson argued that work simulations can be classified according to the extent that they involve hands-on performance and are performed in a real-world setting. Palmer, Boyles, Veres, and Hill (1992) described work simulations for clerical jobs that involved proofreading and filing. These work simulations were clearly hands-on and the tasks performed were very close to a real-world setting and thus could be considered to have high fidelity. At the other extreme, situational judgment tests (Weekley & Ployhart, 2006) present job-related problem scenarios and ask respondents about various actions that might be taken in response to the scenario. Motowidlo, Dunnette, and Carter (1990) called these tests *low-fidelity simulations*. These work simulations do not require hands-on performance and, rather than experience a real-world problem, examinees read a paragraph or view a video about a hypothetical scenario. A common assessment center simulation, the *in-basket*, falls in the mid-range of fidelity. In an in-basket simulation, the respondent is told to assume a manager's role in an organization and is asked to respond to e-mails, requests, and problem situations (e.g., personnel issues). Although the job might require an incumbent to process material in an in-basket, the material to be processed is unlikely to be identical to the information found on the job, and the time limits and other constraints of the in-basket simulations reduce the fidelity of the simulation.

Bandwidth concerns the breadth of the work simulation. The Palmer et al. (1992) proofreading and filing work simulations covered important aspects of the job but did not cover all of the job. Clerical employees engage in various forms of communication (e.g., face to face and e-mail) and maintain interpersonal relationships. Felker et al. (2007) noted that, ideally, all job tasks under all important working conditions would be incorporated into a simulation, but there are several factors that prevent this. First, there are feasibility issues. Simulating all tasks in a job would usually result in a simulation that is too long and too expensive to build and administer. Also, a work simulation should not include tasks that might result in damage to expensive equipment or that might result in injury. Second, job tasks vary widely in their duration, frequency, difficulty, and importance. Although clerical employees may sharpen pencils, it is probably not a task that one would want to include in a work simulation because of its trivial nature. Later in the chapter, we discuss approaches to selecting the content of a work simulation.

Work simulations can be distinguished by whether they assess work *processes* or work *products* (Felker et al., 2007). When a work product is important but is not how one arrives at the end result, the work simulation and its scoring should focus on the output of the task. For example, a simulation for a clerical employee might be preparing a statistical table. The preparation of such a table can be approached in different ways. The work process is less important than producing an accurate and presentable table. On the other hand, the focus of the work simulation may be on the process when the performance of particular steps is important. For example, a clerical employee who needs to respond to an angry customer needs to communicate politely, correctly receive and convey information, and diffuse the customer's anger. In such a simulation, the process by which the employee handles the interaction is important. As another example, consider a work simulation test for a plumber fixing a clogged sink drain. There are multiple ways to score such a simulation,

two of which are a checklist and an outcome-based measure. When a plumber unclogs a drain, there are multiple steps to accomplish the task. If an evaluator had a checklist and watched the applicant go through these steps, the evaluator could check off the items as a way to score the process. Alternatively, imagine the same situation, but instead of completing a checklist, the evaluator leaves and comes back in fifteen minutes and simply checks to see if the drain is unclogged—a simple outcome-based scoring method where the end result is either achieved or not.

Testing format is another way in which work samples may vary. Often, process steps may be scored "go/no go" (i.e., either the person performed the step or not). On the other hand, situational judgment tests are often scored by determining how many optimal responses examinees selected when they were provided a set of response options and they could indicate the most and/or least likely action they would take in a given situation. Simulations with higher fidelity (e.g., those assessing aircraft maneuvers) involve machine scoring so that variables such as reaction time can be assessed. Assessment centers often require raters to judge performance using behaviorally based rating scales.

This chapter will focus on work simulations and the use of work analysis in developing such measures. Work simulations are useful for personnel selection, criterion development in employment test validation studies, job training, licensure examinations, training certifications, and in various education applications (Ferrari, Taylor, & VanLehn, 1999).

THE ROLE OF WORK ANALYSIS IN WORK SIMULATIONS

Work analysis provides the foundation for developing work simulations. To develop a realistic work simulation, one must understand the nature of the work being simulated via a work analysis. This section describes issues in the conduct of work analysis that may be considered when developing work simulations. Later in this chapter, we describe methods for using work analysis information to create work simulations.

The first issue to consider is the level of specificity of the work analysis. There is a continuum of specificity ranging from the description of high-level job attributes to the description of minute details of a job. At one end of the continuum, competency modeling involves the use of broad descriptors of human attributes (e.g., problem solving) that make "an effort to understand the organization's mission, values, strategy, and broad goals" (Sanchez & Levine, 2001, p. 84). This level of analysis may be too general for the purpose of creating work simulations. On the other hand, the study of individual movements, characteristic of time and motion studies (Taylor, 1911), may be too specific. Between these two extremes, using traditional job analysis, one identifies duties, tasks, and the knowledge, skills, abilities and other characteristics (KSAOs) needed to perform the tasks. This is probably most useful for work simulations because the tasks often form the basis for the simulation to be created. Using the linkages between tasks and KSAOs, one can identify specific KSAOs measured by the work simulation.

A related issue to consider is the source of work analysis data. Peterson and Jeanneret (2007) made a distinction between inductive and deductive job analysis. Deductive methods are those "that emphasize the use of existing knowledge or taxonomies of job information during analysis of the focal job" (Peterson & Jeanneret, 2007, p. 13). Examples include the Position Analysis Questionnaire and the Occupational Information Network (O*NET). Inductive methods, on the other hand, involve collecting new, detailed information, usually from subject matter experts (SMEs). Inductive approaches are typically most useful as input for work simulations because the level of detail provided by SMEs is useful for creating simulation exercises.

Once a work analysis has been conducted, one must consider the purpose and context in which work simulations are developed. The purpose for creating a work simulation will inform many

decisions made during its development. The advantages and disadvantages of work simulation are described below.

Advantages and Limitations of Work Simulations

Work simulations are used increasingly because of their many advantages compared with traditional methods of gauging future job performance quality (Ames & Bailey, 2005; Felker et al., 2007). These advantages include practicality, useful levels of criterion-related validity, perceptions of less adverse impact than other selection procedures, and positive applicant reactions (Callinan & Robertson, 2000; Schmidt & Hunter, 1998; Truxillo et al., 2004). Some employers turn to work samples because they perceive that the tests are less likely to be challenged as part of legal discrimination cases.

However, work simulations have several disadvantages (Callinan & Robertson, 2000; Felker et al., 2007; Truxillo et al., 2004). Work simulations can be expensive to build and maintain. The development process typically involves substantial input from individuals who are knowledgeable about the job. Also, test development expertise is required to assess the job content to be simulated and to determine how performance will be scored reliably. Some work simulations may require expensive equipment to be built or adapted. Because work samples are targeted to a job, an employer with many types of jobs may need many work simulations. Also, as jobs change, the work simulation may become outdated and require revision. Work simulations can be costly to administer. Often, they need to be administered individually or with a small number of applicants, and test administrators need to be trained. Finally, some work simulations require that respondents have substantial knowledge of the job. Such work simulations may not be appropriate for entry-level selection if the respondents are to receive extensive training after being hired. For example, because new police cadets are taught how to fire a gun after they are hired (during academy training), a work simulation used to screen police officer applicants may not involve firing weapons.

PSYCHOMETRIC CHARACTERISTICS OF WORK SIMULATIONS

In the field of industrial/organizational psychology, there is some controversy over the use of simulations for selection. For selection, several low-fidelity simulations (e.g., situational judgment tests and situational interviews) measure several different constructs simultaneously, and it is difficult to assess internal consistency, reliability, and construct validity. Therefore, these kinds of measures are considered methods of measurement rather than measures of a single construct. Below, we discuss the reliability of high-fidelity work samples.

Reliability

Work simulations in military settings have reported very high interrater agreement (Felker et al., 2007; Knapp & Campbell, 1993). For example, Carey (1990) and Felker et al. (1988) reported agreements exceeding 90% between test scorers and "shadow" scorers for a variety of Marine Corps job sample tests. Further, Hedge, Lipscomb, and Teachout (1988) reported agreements across pairs, ranging from approximately 75% to 90% across teams of test administrators for three Air Force occupations. Felker et al. (2007) noted that a plausible explanation for these high reliabilities is the care with which the work sample tests were developed. During the Joint Performance Measurement project, the military services devoted substantial effort and resources to the design and administration of work samples (Campbell et al., 1990; Green & Wigdor, 1991). As such, these simulations were subject to extensive pilot testing and were revised as needed to achieve high reliability.

One might assume that work samples developed as either predictors or as criterion instruments in the published literature reflect the high end of care taken to develop such measures. Often, work

simulations in applied settings may be developed with far fewer resources and may yield much lower reliability. Work simulations can be expected to have lower reliability when the measures are developed quickly and are not pilot tested; when the scoring requires subjectivity, lacks standardization in administration, and has too few scorable tasks; and when scorers are untrained.

Validity

Much research suggests that work simulations are highly valid predictors of job performance (Hunter & Hunter, 1984; Reilly & Warech, 1993; Schmitt, Gooding, Noe, & Kirsch, 1984). For example, Hunter and Hunter (1984) obtained a mean corrected validity coefficient of .54 across studies assessing the validity of work simulations, and Schmitt et al. (1984) found an observed validity of .32 (albeit on a fairly small sample).

More recently, Roth et al. (2005) provided a comprehensive summary of the criterion-related validity of work simulations. They found that previous meta-analyses were flawed in their lack of data or there were methodological problems (e.g., range enhancement of validities by including only those who scored at the top and bottom thirds of the distribution). Their results show that work simulations had a mean observed validity of .26 (k = 54, n = 10,469), which increased to .33 when measures of job performance (supervisor ratings) were corrected for attenuation. As previously mentioned, it is important to remember that the general category of work simulations encompasses the measurement of multiple constructs (as opposed to measures that assess a single construct, such as general mental ability). The validity of a work sample will likely vary with the constructs assessed.

There are several constructs that potentially moderate the validity of work samples, such as the status of the participant (applicant vs. incumbent), measure of job performance (subjective vs. objective), criterion versus predictor conceptualization, type of sample (military vs. nonmilitary), and job complexity (Roth et al., 2005). Their results were inconclusive regarding sample differences (applicant vs. incumbent) because of insufficient data for applicants (k = 1, n = 24). The use of objective versus subjective measures did not have a strong influence on sample validity (.27 vs. .26, respectively). Whether the work simulation was used as a predictor or as a criterion also was not a meaningful moderator, as the correlations were .29 and :25, respectively. The use of military versus nonmilitary samples did not seem to moderate the validity (.25 vs. .28, respectively), nor did job complexity (correlations ranged from .25 to .28 for each of three levels of job complexity). Roth et al.'s (2005) results suggest that work sample measures may not be as valid as previously believed and that meaningful moderators have yet to be discovered.

The criterion-related validity of low-fidelity work simulations, in the form of situational judgment tests, has been evaluated in many primary studies (Chan & Schmitt, 1997; Hanson & Borman, 1989; Motowidlo et al., 1990; Smith & McDaniel, 1998). Two meta-analyses have examined the criterion-related validity of situational judgment tests (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel et al., 2007). In the second and more recent meta-analysis, the overall validity of situational judgment tests across 118 coefficients was .26 (n = 24,756), regardless of instruction type. These validity results are almost entirely based on concurrent validity studies (e.g., research typically conducted using job incumbents, rather than applicants, as subjects). Their results showed that response instructions influenced the constructs measured by the tests, such that tests with knowledge instructions had higher correlations with cognitive ability and tests with behavioral tendency instructions showed higher correlations with personality constructs. Results also showed that response instructions had little moderating effect on criterion-related validity.

Regarding the validity of assessment centers, there is substantial agreement that these measures can yield scores that predict job performance (Dean, Roth, & Bobko, 2008), and meta-analyses have provided corrected validities of .28 (Hardison & Sackett, 2004) to .36 (Gaugler, Rosenthal,

Thornton, & Bentson, 1987). In fact, some researchers have stated that the "predictive validity of assessment centers is now largely assumed" (Borman, Hanson, & Hedge, 1997, p. 313).

In the educational arena, research and thought have been devoted to considerations of aspects of validity that guide the development of work simulations. Miller and Linn (2000) outline six aspects of construct validity of performance-based assessments. These include content, substantive, structural, generalizability, external, and consequential. The content aspect of validity focuses on the relevance and representativeness of the assessment's content. The substantive aspect focuses on processes used by examinees when they respond and on the similarity of the processes with the construct the assessment is designed to measure. The structural aspect addresses the adequacy and appropriateness of scoring and scaling. The generalizability aspect focuses on replicability of results across various levels and facets of the assessment procedure (e.g., across raters and tasks). The external aspect concerns convergent and discriminant evidence showing a relationship (or non-relationship) between the assessment and other performance measures. The consequential aspect focuses on the degree to which assessments have the intended positive effects and plausible unintended negative effects. These aspects of validity are important to consider during the development of work simulations.

Subgroup Differences

Much of the literature surrounding subgroup differences and work simulations suggest that these tests are associated with lower levels of adverse impact than traditional measures of cognitive ability (Cascio, 2003; Gatewood & Field, 2001; Reilly & Warech, 1993; Salgado, Viswesvaran, & Ones, 2001; Schmitt, Clause, & Pulakos, 1996). Bobko et al. (2005) point out that much of that literature has used incumbent samples rather than applicant samples and, as a result, the estimates are subject to range restriction from prior selection. Thus, these estimates would likely underestimate the magnitude of subgroup differences.

Concerning subgroup differences on assessment centers, Dean et al. (2008) conducted a meta-analysis and found a mean d for Black–White differences of .52 (k = 17, n = 8,210). When examinee type (applicant vs. incumbent) was examined, d values were .56 for applicants (k = 10, n = 3,682) and .32 (k = 6, n = 1,689) for incumbents. In addition, they found an overall male-female mean d of −0.19, showing that women, on average, obtained slightly higher scores than men.

Bobko et al. (2005) conducted a primary study in which they analyzed two data sets from public sector jobs and showed that the adverse impact of work sample tests may be more extensive than previously thought. Two different assessment centers were used, one for each job. The first assessment center included a technical exercise, an in basket, a counseling exercise, and a set of oral responses to incidents (interruptions to other work). The exercises were scored for content (what was said or written) and for process (oral or written communication skills). The Black–White d values ranged from −0.06 for the content of the counseling exercise to 0.80 for the technical score (the d on the overall score was 0.73). Positive values indicate White subjects, on average, scored higher than Black subjects. Bobko et al. noted that the process dimensions were associated with smaller Black–White d values than other components. The second assessment center included three exercises: a map reading test, a technical exercise, and a role-play exercise. The Black–White d values ranged from 0.12 for the human relations ability score on the role play to 0.80 for the map-reading score (the d on the overall score was 0.73). The three dimensions measured by the role-play exercise were associated with the smallest Black–White differences. Note that, in both cases, the numbers of Blacks tested were small (n = 31 and 33, respectively), but the results were consistent. The authors note that the values of d for both work sample exams are close to the value of d (0.72) associated with the Wonderlic test of general mental ability when selecting applicants for medium complexity jobs (Roth, Bevier, Bobko, Switzer, & Tyler 2001). For both data sets, to the extent that the exercises

were *g*-loaded (e.g., technical exercise and map reading), the *d* values are higher than exercises loaded with more personality-related constructs.

Regarding subgroups differences on situational judgment tests, Whetzel, McDaniel, and Nguyen (2008) conducted a meta-analysis of mean race and sex differences in situational judgment test performance. On average, White subjects performed better on situational judgment tests than Black, Hispanic, and Asian subjects (d =0.38, k = 62, n = 42,178; d = 0.24, k = 43, n = 14,195; d = 0.29, k = 25, n = 16,515, respectively). Women performed slightly better than men (d = −0.11, k = 63, n = 37,829). They investigated two moderators of these differences: (a) loading of *g* or personality, and (b) response instructions. Mean race differences between Black, Hispanic, Asian, and White examinees in situational judgment test performance were largely explained by the cognitive loading of the situational judgment test such that the larger the cognitive load, the larger the mean racial differences. The effect of personality loading on race differences (Black–White and Asian–White) were smaller to the extent that situational judgment tests are correlated with emotional stability. Hispanic–White differences were smaller to the extent that situational judgment tests were correlated with conscientiousness and agreeableness. Cognitive loading had minimal effect on male-female score differences; however, score differences were larger, favoring women, when situational judgment tests were correlated with conscientiousness and agreeableness. Knowledge response instructions showed greater race differences than behavioral tendency instructions. The mean correlations show that these differences are largely due to the greater *g*-loading of knowledge instructions.

In summary, work simulation measures, whether assessment center exercises or situational judgment tests, exhibit nontrivial performance differences based on race, such that, on average, White subjects perform better than Black subjects on these measures. For both assessment centers and situational judgment tests, this is more likely if the tests are *g* loaded rather than personality (noncognitively) loaded. In both assessment centers and situational judgment tests, women, on average, performed slightly better than men, perhaps because of the interpersonal nature of both kinds of assessments.

As a result of these findings, one might consider developing video-based situational judgment tests to reduce the *g*-loading of such assessments. For assessment centers, rather than have people write their answers to exercises (e.g., the in-basket), they might describe their answers verbally to scorers. Assessment centers also might focus more on role playing rather than exercises that require a high degree of analytical thinking. However, given the useful validities of *g*-loaded measures, one may achieve lower group differences at the cost of reduced validity (Ployhart & Holtz, 2008).

HOW TO CREATE WORK SIMULATION MEASURES

Although it is possible to purchase work simulations off the shelf, for most purposes it is more common to develop them for a specific job or organization because there often are job knowledge requirements embedded in the simulation. In this section, we describe methods for developing these tests. Because performance tests come in a variety of types and formats (e.g., assessment centers, situational judgment tests, hands-on performance tests), we provide a generic approach based on what these measure have in common. Although the approach described is generic, we use examples that focus on one or another specific kind of test. When needed, we discuss the development of high- and low-fidelity simulations separately. For readers interested in additional guidance, we recommend Gatewood and Field (2001), Guion (1998), and Felker et al. (2007). This part of the chapter will follow an outline similar to that used by Felker et al. (2007) and Truxillo et al. (2004).

How to Select Test Content

When constructing a work simulation measure, the first decision concerns the part(s) of the job domain to be tested. Ideally, all tasks would be tested under all important working conditions;

however, for safety and other practical reasons, this is usually not feasible. Similar to the development of other types of tests, the challenge is to select a small number of tasks that represent the larger pool of job tasks so that test performance can be generalized to the job. This is typically conducted in two steps: (a) specify the total performance domain for the job, and (b) devise a valid and defensible sampling strategy for selecting tasks from that domain.

The job performance domain can be described in a number of ways, either through a complete job analysis describing tasks and KSAOs that compose a job, a critical incident approach in which SMEs describe examples of performance at various levels of proficiency, or a competency-based approach in which SMEs describe, in broad strokes, what attributes are needed to perform a job. Competency-based approaches often do not provide sufficient detail for creating work simulations, and they will not be described further here. For more description of competency-based approaches, see Shippman et al. (2000) and Ulrich, Brockbank, Yueng, and Lake (1995).

The use of a particular job analysis method depends on the type of simulation to be developed. Job and task analyses are frequently used to develop higher-fidelity simulations (e.g., hands-on measures of performance and assessment centers), whereas critical incidents are often used to develop lower-fidelity simulations (e.g., situational judgment tests) because task information provides detailed data useful for developing assessment center exercises; critical incidents provide big picture situations that are useful for creating scenarios for situational judgment tests. Strategies for sampling from the domain vary based on whether high- or low-fidelity simulations are developed, and they are described separately below.

High-Fidelity Performance Tests

In addition to a complete list of tasks, when higher-fidelity tests are being constructed the results of a job and task analysis should focus on contextual or environmental conditions of task performance, chronological dependences of task performance (which tasks require performance before other tasks), and interaction requirements (when the performance of tasks requires interaction with others).

Felker et al. (2007) described two methods of task selection commonly used by the U.S. military: (a) the four- and eight-factor models and (b) the Difficulty, Importance, and Frequency model. In the first method, test developers rate each task in the domain on up to eight of the following descriptors:

1. Percentage of the workforce performing the task
2. Task delay tolerance (degree of flexibility before task must be performed)
3. Consequences of inadequate task performance
4. Task learning difficulty
5. Percentage of time spent performing task
6. Probability of deficient task performance
7. Immediacy of task performance (urgency)
8. Frequency of task performance

The first four of these descriptors comprise the four-factor model and the entire set make up the eight-factor model. Job analysts/SMEs rate each task in the domain on each of the four or eight descriptors, and cutoffs are set for defining the relevance of each factor. Examples of cutoffs are the percentage of the workforce performing the task (cutoff is 40% or more) and consequences of inadequate performance (1 = minor, 7 = major; cutoff is 5). Tasks that meet or exceed a selected set of cutoffs are selected for testing.

The Difficulty, Importance, and Frequency model also requires that SMEs provide task ratings. Using this method, job analysts/SMEs rate each task on the following:

1. *Difficulty*: Learning difficulty, probability of deficient performance, or both
2. *Importance*: Consequence of inadequate performance, task delay tolerance, and time spent performing the task
3. *Frequency*: How often the task is performed and the percentage of people performing the task

As with the four- and eight-factor models, SMEs provide ratings on each of the Difficulty, Importance, and Frequency dimensions, and decisions are made about whether the tasks should be included for testing.

Gatewood and Field (2001) described other important criteria to be used in selecting tasks for testing, including the following:

1. Tasks in which the total time required for completion is reasonable
2. Tasks that are representative of the job in terms of difficulty or complexity, as tasks that are too easy or too difficult will not help to distinguish among examinees in terms of proficiency
3. Tasks that require less expensive materials, equipment, or facilities
4. Tasks that have standardized operations or products or have easily defined verbal or interaction components, as it is easier to develop and score situations based on such tasks

Once tasks are generated, most job analysis processes involve the specification of KSAOs for performing those tasks. As indicated by Harvey, Anderson, Baranowski, and Morath (2007), knowledge refers to specific types of information that people must know to perform a job (Williams & Crafts, 1997); skills can be thought of as the capabilities needed to perform a task, which can be developed over time and with exposure to multiple situations; and abilities are an individual's relatively enduring capabilities for performing a particular range of different tasks (Fleishman, Costanza, & Marshall-Mies, 1997). Other characteristics include occupational values and interests and work styles (Peterson, Mumford, Borman, Jeanneret, & Fleishman, 1997); personal preferences and interests (Holland, 1973); and individual difference variables (Jackson, 1967) that facilitate the performance of a job.

This discussion so far has focused primarily on hands-on performance tests; however, the processes described above are useful for developing assessment centers as well. As Tsacoumis (2007) noted, the first step in developing an assessment center is to review current job analysis data. Specifically, the exercise developer needs to review the task-KSAO linkages (which KSAOs are needed to perform which tasks) to generate ideas about possible exercises that simulate job tasks.

Regardless of which kind of high-fidelity simulation is being developed, there are a variety of issues to consider during task selection. Not only should characteristics of the tasks themselves be considered (e.g., importance, difficulty, and linkage to KSAOs), but the practical issues surrounding the testing of those tasks (e.g., equipment availability) are of paramount importance.

Lower-Fidelity Performance Tests

As mentioned previously, lower-fidelity performance tests (e.g., situational judgment tests and oral interviews) present applicants with problem situations that may be encountered on the job and ask them to describe, verbally or in writing, how they would respond.

Domain sampling strategies for these kinds of tests typically involve having SMEs generate critical incidents. Critical incidents (Flanagan, 1954) include descriptions of a situation encountered by

an employee, what the employee did in response to the situation, and the result of the employee's actions. A form for collecting critical incidents is shown in Figure 22.1. To create a situational judgment test, hundreds of such incidents are collected, typically as part of a series of workshops; they are then sorted by job analysts or SMEs into dimensions according to behavioral similarity. Once performance dimensions are identified, a second group of SMEs engages in a retranslation process in which they sort incidents into the given dimensions. Once agreement is reached about which incidents describe each dimension, the test development process can be started.

How to Develop Work Simulations

The test development process for both higher- and lower-fidelity work simulations entails generating test situations that represent each of the tasks or performance dimensions identified above. For higher-fidelity simulations, in which the respondent actually performs a task or part of a task, the goal is to generate a number of task situations. For lower-fidelity simulations, in which respondents are asked to describe what they would do or effective actions in a situation, the goal is to generate a

Critical incident form	Participant #_____
1. What was the situation leading up to the event? [Describe the context.]	
2. What did the employee do?	
3. What was the outcome or result of the employee's action?	
4. Circle the number below that best reflects the level of performance that this event exemplifies.	
1 2 3 4 5 6 7	
Highly ineffective / Moderately effective / Highly effective	

Figure 22.1 Sample critical incident form.

number of problem situations. In this section, we describe general approaches for developing and scoring high- and low-fidelity simulations. We then discuss procedures for setting passing scores.

High-Fidelity Work Simulation Development

Once the tasks are selected, there are many other issues to consider when developing a work sample. High-fidelity performance tests require examinees to perform a task, or an essential part of a task, under conditions similar to how the task is performed on the job. As with all kinds of tests, work samples must be administered in the same way and under the same conditions so that all examinees have the same opportunity to demonstrate their ability. Thus, instructions for administering work sample tests are critical for standardizing how the tests are administered and scored.

Because high-fidelity performance tests represent abstractions of the actual work performed on the job, decisions need to be made regarding which compromises in psychological fidelity are necessary to accommodate the practical constraints of the test environment (Felker et al., 2007). Physical fidelity often must be sacrificed when it is impossible to duplicate exact working conditions or when doing so is costly. Instead, these conditions are simulated in such a way as to elicit the same knowledge and skills needed to perform the task, thus ensuring the test's psychological fidelity. For example, it is generally not feasible to test all truck drivers on all possible driving conditions in a standardized way. However, requiring examinees to drive around portable obstacles at certain speeds and brake at different speeds (as determined by the job analysis) may adequately simulate skills needed in many driving conditions, such as wet roads and heavy traffic (Felker et al., 2007).

When administering work simulation tests, some authenticity may be lost because of practical considerations. For example, tasks may be tested only in part because they are too long or too trivial to be tested in their entirety. In such cases, examinees might be given the information necessary to perform the task or they might be asked to walk through a particular step rather than demonstrate it when actually testing the step is infeasible (Hedge & Teachout, 1992).

The next step in developing high-fidelity performance tests is to devise procedures for scoring performance. Because scoring high-fidelity simulations is often done in real time as the task is performed, scoring requires familiarity with the task. Felker et al. (2007) make several recommendations for ensuring that the performance of steps of a task are accurately and reliably scored. First, performance steps need to be observable. Thus, if a checklist is used, similar to that in Figure 22.2, action verbs should be used (e.g., set the X switch, install the thingamajig). Steps that call for "checking," "inspecting," "reading," or "observing" are not observable in that the scorer does not know if the examinee is really "checking" or "reading," and if they are, one does not necessarily know what they are "checking." Second, standards for performance should be objective. For example, "drives an 18-wheel vehicle around a set of obstacles adequately" is not particularly useful because there is no observable method for assessing adequacy. "Drives an 18-wheel vehicle around a set of obstacles at 35 mph without knocking down any cones" specifies the standard to be met by the examinee. Another important recommendation is to use job aids in the tests, especially when the tasks are lengthy or complicated or when job incumbents are not expected to memorize specific procedures. Having them look something up in a manual can be incorporated into the test and still be scorable. Training test administrators is an important feature of work sample tests. Training approaches should focus on scorers' ability to observe behavior and to make judgments based on their observations (Hedge & Kavanagh, 1988). In addition, having scorers practice performing, observing, and rating the tasks they will be scoring is an important training component. Using two or more independent raters is helpful to identify possible scoring differences and to assess interrater reliability.

For scoring product tests, one can develop a score sheet that documents what the end-product should look like. It also is helpful to develop a sample of the work product so that scorers can compare the examinee's product to the correct work product.

Unclogging a sink drain

Test date: _____ Examinee: _____

Scorer: _____ Examinee ID: _____

Say: This sink is clogged and you need to enable water to run through it smoothly. You are to use the materials provided. Do you have any questions? Begin.

Performance steps:	Go	No-Go
1. Remove the basket strainer.	___	___
2. Run hot water until it reaches two inches deep.	___	___
3. Try a suction plunger.	___	___
4. Put a pail under the sink.	___	___
5. Remove the cleanout plug and washer.	___	___
6. Try to unclog with screwdriver.	___	___
7. Use an auger.	___	___
8. Run hot water to clear residue when unclogged.	___	___

Figure 22.2 Example of a score sheet for a plumber task. (Adapted from http://www.atexinspects.com/Plumbing-Clogs.html.)

Assessment center tasks (e.g., in baskets, analysis exercises, and role plays) are high-fidelity simulations often used for supervisory or managerial positions (Tsacoumis, 2007). As described above, for other high-fidelity simulations, one defines the job content using a job/task analysis and identifies the KSAOs or competencies to be assessed. Then, working with SMEs (who are approximately one level above the target position), the developers generate ideas for different job simulations. One simulation might be a role play in which the examinee, in the role of a supervisor, meets with an employee (the scorer) to discuss a project. Another simulation might be an in-basket task during which examinees review materials typically found on a supervisor's desk (e.g., incorrectly completed time sheets, scheduling conflicts, and employee requests), and the examinee indicates the action he or she would take for each item. An analysis exercise might include some technical feature of the job that a supervisor might need to assess and present to employees. Then, rating scales are developed that provide scorers with observable behaviors at various levels of proficiency against which to compare examinee behavior. An example of such a rating scale is provided in Figure 22.3. Specific details about the development of each kind of assessment center exercise are beyond the scope of this chapter. For a description of how to develop and score such exercises, see Tsacoumis (2007).

Low-Fidelity Work Simulation Development

As mentioned previously, low-fidelity performance tests present examinees with problem situations that might be encountered on the job and ask them to provide information about a set of possible responses. Critical incidents, as described above, are useful for this purpose. Once incidents are generated and sorted into dimensions, scenarios are created that form the question in a situational judgment test. When the questions are developed, the response options are generated. One approach for generating response options is to ask the group of SMEs who generated the critical incidents to describe how outstanding, average, and poor job incumbents would deal with each situation. Another approach is to ask job incumbents with low levels of experience how they would deal with each problem (Motowidlo, Hanson & Crafts, 1997).

Role Play Dimension: Plan, Organize, and Prioritize		
Plan, Organize, and Prioritize involves the ability to plan activities to reach a goal and to organize information. It also involves managing multiple activities simultaneously to decide their priority toward accomplishing a goal.		
1 2 3	4	5 6 7
Ineffective • Rarely matched assignments to capabilities or developmental opportunities (did not see that Pat was capable of X activity and assigned it to Alex) • Made work unit plans but ignored customer service requirement • Overloaded high-performing employees by giving them too much work (gave competent employee, Tyler, more work than could be accomplished in time period)	Effective • Scheduled and staffed appropriately according to workload individual capabilities • Determined the steps that needed to be taken to meet objectives • Gave appropriate amount of work to people at all levels of proficiency	Very effective • Organized and planned ahead about the resources needed for the day; looked at schedules and data to meet work unit goals • Effectively created short-term and long-term work unit plans using historical data and workload projections • Provided advance notification of workload fluctuations to other areas of the organization, offering own resources to ensure organizational goals are met (suggested that Lee work in another department to help them finish deliverables)

Figure 22.3 Sample assessment center rating scale.

The scoring algorithm is developed by having SMEs rate the responses for effectiveness. The mean of their ratings and agreement indices are then used to scale the response options. When asking examinees to rate the effectiveness of each option, their responses are scored against the SMEs' ratings, and higher scores are achieved to the extent that they are in agreement. When asking examinees to select what they would most or least likely do, the SMEs' ratings would be used to identify the most or least effective response, and applicants correctly choosing each alternative would receive a score of 1 for the item; otherwise they would receive a 0. Other scoring methods are described by McDaniel and Whetzel (2007) and Motowidlo et al. (1990). An example of a situational judgment test item is shown in Figure 22.4.

Setting Passing Scores

Many researchers (e.g., Schmidt, Mack, & Hunter, 1984) have shown that top-down approaches to selection enhance the utility of a selection procedure for organizations, and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society of Industrial and Organizational Psychology, 2003) noted that, with valid predictors, a top-down approach is best for organizations. However, because this practice maximizes adverse impact (especially when tests are *g*-loaded), users of tests often have used cutoff scores such that everyone above a certain score is equally likely to be selected. Consequently, the process used for setting defensible cutoff scores is of great importance, especially if the tests are used for high-stakes decisions. If it is necessary to set a passing score for a work simulation, particularly one that reflects minimum job performance, a commonly used method, based on expert judgment, is the Angoff method (Angoff, 1971). The Angoff method is typically used for content validated tests, and it has been used to set passing scores for work samples (e.g., Truxillo, Donahue, & Sulzer, 1996) and certification tests (e.g., Busch & Jaeger, 1990). Despite its subjectivity, the Angoff method has withstood legal challenge (e.g., Biddle, 1993).

You are working on a project with a coworker. This project has a very tight time frame and your boss is very interested in having it completed on time. Your coworker is not pulling his/her share of the workload.

a. Confront your coworker with his/her nonperformance.
b. Leave the coworker alone hoping that the problem will correct itself when he/she sees how much work you are doing.
c. Talk to your boss about the situation so that he will intervene.
d. Talk to another coworker about this problem to see if this has happened before.
e. Ask your boss for a time extension on the project.
f. Ask your coworker if he/she is having family problems that could be affecting his/her work.

Figure 22.4 Example of a situational judgment test item.

Using the Angoff method, judges are asked to review test items and to estimate the percentage of minimally competent persons who could answer the item correctly or the likelihood that a minimally competent person would answer the item correctly. The cutoff is based on the average estimate across items and judges. Detailed recommendations regarding the implementation of this method are given in several reviews (e.g., Biddle, 1993; Truxillo et al., 1996). These involve using a fairly large number of judges (7–10) who represent demographic and geographic diversity and various organizational units so that the resulting cutoffs will seem fair to stakeholders.

Extensions of the Angoff approach include having panelists estimate the typical score that an examinee will earn on a question (Hambleton, Jaeger, Plake, & Mills, 2000), having panelists not only estimate the minimum number of score points for borderline examinees, but also estimate the distribution of scores of borderline examinees at basic, proficient, and advanced proficiency levels (Reckase, 2000).

FUTURE POTENTIAL DEVELOPMENTS

Since the mid-2000s, technological advances have greatly changed the field of testing, and the implications for simulation testing are profound. These changes involve how simulations are both administered and scored. There are also advances in job analysis, although at a less dramatic pace. As the processes associated with both job analysis and testing evolve, the use of work simulations will continue to advance in terms of delivery and frequency of use.

Concerning the delivery of work simulations, virtual reality is becoming more of an option. The use of avatars (i.e., computer-based personalities that individuals create and manage in an online environment) is prevalent in many areas of life. Even at grocery stores, one can follow the movements of an avatar in self-serve checkout lines. One can imagine the flexibility in using avatars instead of video-based tests for situational judgment tests.

Work simulations may also be used more frequently in on-the-job training. In a variety of work tasks, particularly those involving computer applications, one can be taught how to complete certain tasks using simulations. For example, computer security applications involving fingerprint scanners are often accompanied by tutorials that teach the user how to scan the fingerprint device. Such software can involve practice trials that assess the adequacy of the user's mastery of the presented materials. Work simulations developed in virtual worlds, such as Second Life, have been used to simulate operating rooms (Gerald & Antonacci, 2009). Second Life also has been used to simulate Air Force bases, using the program MyBase, which allows the public to access information about the Air Force using virtual characters who chat by voice and text. Another virtual continent called SciLands is devoted to science and technology education. Although these examples represent high-fidelity simulations built with many bells and whistles, there is some evidence that low-fidelity

simulation can achieve high levels of transfer of training without the costs associated with the use of high-fidelity simulators (Thomas, 2009).

Use of the Internet for testing has been a hotly debated topic. Tippins (2009) summarized several issues with using unproctored Internet testing (e.g., test security and cheating). Because of the low cost of administering tests through the Internet (no travel for examinees, no test administrators at various locations), many companies are using unproctored testing, especially for noncognitive tests. Much attention has been devoted to statistical detection of cheating as well as prevention of cheating (e.g., requiring some form of identification) prior to taking the test. Solutions to these issues include the development of a large item bank of simulations so that all examinees do not receive the same stimulus.

FURTHER READING

The references that follow present additional reading and resources concerning work simulations. Our recommended reading includes major reviews of work simulations with emphasis on applicant screening, work samples (narrowly defined), and situational judgment tests. Referring readers to Internet resources is a tricky business because web site addresses may change. We list sites that are likely to have some permanence. Readers may wish to conduct their own web searches. We also did not provide any references to technology sources (e.g., software) because it is outside of our expertise; the field is evolving rapidly and anything we could offer would likely be out of date by the time this book is published.

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment, 13*(1), 1–10.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8*, 248–260.

Felker, D. B., Curtin, P. J., & Rose, A. M. (2007). Tests of job performance. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 319–348). Mahwah, NJ: Lawrence Erlbaum.

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 235–258). Mahwah, NJ: Lawrence Erlbaum.

Ployhart, R., & Weekley, J. (Eds). (2006). *Situational judgment tests: Theory, measurement, and application.* San Francisco, CA: Jossey-Bass.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*, 1009–1037.

Society for Industrial and Organizational Psychology, Inc. (2011). *Work samples and simulations.* Retrieved from http://www.siop.org/workplace/employment%20testing/samplesandsimulations.aspx
This web site provides a brief overview and a link to other employment testing information.

Tsacoumis, S. (2007). Assessment centers. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 259–292). Mahwah, NJ: Lawrence Erlbaum.

Truxillo, D. M., Donahue, L.M., & Kuang, D. (2004). Work samples, performance tests, and competency testing. In J. C. Thomas & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment* (4th ed, pp. 345–367). Hoboken, NJ: John Wiley and Sons.

Stanford School of Medicine, Center for Immersive and Simulation-Based Learning. (2011). Real training from simulated experiences. Retrieved from http://cisl.stanford.edu[2]

National Training and Simulation Association (homepage). Retrieved from http://www.trainingsystems.org[3]

[2] This organization develops work simulations to aid in the training of medical professionals.

[3] This organization is a professional organization of companies that provides simulations used for training, primarily for the U.S. Military.

REFERENCES

Ames, B., & Bailey, B. (2005, June). *Pennsylvania's computer-administered work simulation assessments.* Paper presented at the 29th annual meetings of the IPMAAC, Orlando, FL.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 121–208). Washington, DC: American Council on Education.

Biddle, R. E. (1993). How to set cutoff scores for knowledge tests used in promotion training, certification, and licensing. *Public Personnel Management, 22,* 63–79.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 25,* 499–513.

Bobko, P., Roth, P. L., & Buster, M. A. (2005). Work sample selection tests and expected reduction in adverse impact: A cautionary note. *International Journal of Selection and Assessment, 13,* 1–10.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48,* 299–337.

Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27,* 145–163.

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248–260.

Campbell, C. H., Ford, P., Rumsey, M. G., Pulakos, E.D., Borman, W.C., Felker D.B., … Riegelhaupt, B. J. (1990). Development of multiple job performance measures in a representative sample of jobs. *Personnel Psychology, 43,* 277–300.

Carey, N. B. (1990). *An assessment of surrogates for hands-on tests: Selection standards and training needs (CRM 90-47).* Alexandra, VA: Center for Naval Analyses.

Cascio, W. (2003). *Managing human resources: Productivity, quality of work life, and profits* (6th ed.). Boston, MA: McGraw-Hill.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in SJTs: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159.

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: A meta-analysis. *Journal of Applied Psychology, 93,* 685–691.

Employment Technologies Corporation. (2008). Easy Simulation Teller Vision. Retrieved from http://www.etc-easy.com/_products/tellervision/index.htm

Felker, D. B., Crafts, J. L., Rose, A. M., Harnest, C. W., Edwards, D. S., Bowler, E. C., … McHenry, J. J. (1988). *Developing job performance tests for the United States Marine Corps infantry occupational field* (AIR-47500-9/88-FR). Washington, DC: American Institutes for Research.

Felker, D. B., Curtin, P. J., & Rose, A. M. (2007). Test of job performance. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 319–348). Mahwah, NJ: Lawrence Erlbaum.

Ferrari, M., Taylor, R., & VanLehn, K. (1999). Adapting work simulations for schools. *Journal of Educational Computing Research, 21,* 25–53.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 41,* 237–358.

Fleishman, E. A., Costanza, D. C., & Marshall-Mies, J. C. (1997). Abilities. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *O*NET: An occupational information network* (pp. 175–195). Washington, DC: American Psychological Association.

Furst Person. (2004). *How will that job candidate manage your calls? Find our before you hire them.* Retrieved from http://www.furstperson.com/furstperson-tools/interactive-simulations/service-brick-mortar/

Gatewood, R., & Field, H. (2001). *Human resource selection* (5th ed.). Fort Worth, TX: Harcourt College Publishers.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C. III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

Gerald, S., & Antonacci, D. M. (2009). Virtual world learning spaces: Developing a Second Life operating room simulation. *EDUCAUSE Quarterly Magazine, 32.* Retrieved from http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/VirtualWorldLearningSpacesDeve/163851

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis-content validity process. In N. Schmitt, W. C. Borman, & Associates (Eds.), *Personnel selection in organizations* (pp. 3–34). San Francisco, CA: Jossey-Bass.

Green, B. F. Jr., & Wigdor, A. K. (1991). Measuring job competency. In A. K. Wigdor & B. F. Green (Eds.), *Performance assessment for the workplace* (pp. 53–74). Washington, DC: National Academies Press.

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

Hambleton, R. K., Jaeger, R. M., Plake, C. M., & Mills, C. (2000). Setting performance standards on complex educational assessment. *Applied Psychological Measurement, 24,* 355–366.

Hanson, M. A., & Borman, W. C. (1989, April). *Development and construct validation of a situational judgment test of supervisory effectiveness for first-line supervisors in the U.S. Army.* Paper presented at the 4th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Hardison, C. M., & Sackett, P. R. (2004, April). *Assessment center criterion-related validity: A meta-analytic update.* Paper presented at the 2004 meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

Harvey, J. L., Anderson, L. E., Baranowski, L. E., & Morath, R. A. (2007). Job analysis: Gathering job-specific information. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 57–95). Mahwah, NJ: Lawrence Erlbaum.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparisons of three methods of performance appraiser training. *Journal of Applied Psychology, 73,* 68–73.

Hedge, J. W., Lipscomb, M. S., & Teachout, M. S. (1988). Work sample testing in the Air Force job performance measurement project. In M. S. Lipscomb & J. W. Hedge (Eds.), *Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel* (AFHRL-RP-87-58). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Hedge, J. W., & Teachout, M. S. (1992). An interview approach to work sample criterion measurement. *Journal of Applied Psychology, 77,* 453–461.

Holland, J. L. (1973). *Making vocational choices: A theory of careers.* Englewood Cliffs, NJ: Prentice Hall.

Houran, J. (2007, November). *Employee screening with job-simulation videos.* Retrieved from http://www.hvs.com/Jump/?aid=3023

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

ICT Results (2008, June). *Holodeck 1.0? Star Trek-style displays make their debut.* Retrieved from http://www.sciencedaily.com/releases/2008/06/080604195058.htm

Jackson, D. N. (1967). *Personality research form manual.* Goshen, NY: Research Psychologists Press.

Knapp D. J., & Campbell, J. P. (1993). *Building a joint service classification research roadmap: Criterion related issues* (AL/HR-TP-1993-0028). Brooks Air Force Base, TX: Armstrong Laboratory, Manpower and Personnel Research Division.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using Situational Judgment Tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 235–258). Mahwah, NJ: Lawrence Erlbaum.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24,* 367–378.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternate selection procedure: The low fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 241–260). Palo Alto, CA: Davies-Black Publishing.

Palmer, C. I., Boyles, W. R., Veres, J. G., & Hill, J. B. (1992). Validation of a clerical test using work samples. *Journal of Business and Psychology, 7,* 239–257.

Peterson, N. G., & Jeanneret, P. R. (2007). Job analysis: Overview and description of deductive methods. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 13–56). Mahwah, NJ: Lawrence Erlbaum.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., & Fleishman, E. A. (Eds.). (1997). *O*NET: An occupational information network.* Washington, DC: American Psychological Association.

Ployhart, R. E. & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: contemporary practice and theory* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Reckase, M. D. (2000). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts conducted by ACT.* Iowa City, IA: ACT.

Reilly, R., & Warech, M. (1993). The validity and fairness of alternatives to cognitive tests. In L. Wing & B. Gifford (Eds.), *Policy issues in employment testing* (pp. 131–224). Boston, MA: Kluwer.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297–330.

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*, 1009–1037.

Salgado, J., Viswesvaran, C., & Ones, D. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. Anderson, D. Ones, H. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, & organizational psychology* (pp. 165–199). London, UK: Sage.

Sanchez, J. I., & Levine, E. L. (2001). The analysis of work in the 20th and 21st centuries. In N. Anderson, D. Ones, H. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, & organizational psychology* (pp. 71–89). London, UK: Sage.

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

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. *Journal of Applied Psychology, 69*, 490–497.

Schmitt, N., Clause, C., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In C. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (Vol. 11, pp. 115–139). New York, NY: John Wiley.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1983 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422.

Shippman, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., … Sanchez, J. I. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703–740.

Smith, K. C., & McDaniel, M. A. (1998, April). *Criterion and construct validity evidence for a situational judgment measure.* Paper presented at the 13th annual conference of the Society for Industrial and Organizational Psychology, Inc., Dallas, TX.

Society of Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). College Park, MD: Author.

Taylor, F. W. (1911). *The principles of scientific management.* New York, NY: Harper and Brothers.

Thomas, M. (2009). Integrating low-fidelity desktop scenarios into the high fidelity simulation curriculum in medicine and aviation. Retrieved from http://www.unisanet.unisa.edu.au/staff/MatthewThomas/Paper/Thomas_DesktopScenarios.pdf

Tippins, N. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2*(1), 2–10.

Truxillo, D. M., Donahue, L. M., & Kuang, D. (2004). Work samples, performance tests, and competency testing. In J. C. Thomas, & M. Hersen (Eds.), *Comprehensive handbook of psychological assessment* (4th ed., pp. 345–367). Hoboken, NJ: John Wiley and Sons.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issues, illustrations, and recommendations. *Human Performance, 9*, 275–295.

Tsacoumis, S. (2007). Assessment centers. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 259–292). Mahwah, NJ: Lawrence Erlbaum.

Ulrich, D., Brockbank, W., Yueng, A. K., & Lake, D. G. (1995). Human resource competencies: An empirical assessment. *Human Resource Management, 34*, 473–495.

Weekley, J. A., & Ployhart, R. E. (2006). An introduction to situational judgment testing. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 1–10). Mahwah, NJ: Lawrence Erlbaum Associates.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*, 291–309.

Williams, K. M., & Crafts, J. L. (1997). Inductive job analysis: The job/task inventory method. In D. L. Whetzel & G.R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 51–87). Palo Alto, CA: Consulting Psychologists Press, Inc.