Simulations represent more or less exact replicas of tasks, knowledge, skills, and abilities required in actual work behavior. This chapter reviews research on the more traditional high-fidelity simulations (i.e., assessment centers and work samples) and contrasts it with the growing body of research on low-fidelity simulations (i.e., situational judgment tests). Both types of simulations are compared in terms of the following five statements: “The use of simulations enables organizations to make predictions about a broader array of KSAsOs,” “We don’t know what simulations exactly measure,” “When organizations use simulations, the adverse impact of their selection system will be reduced,” “Simulations are less fakable than personality inventories,” and “Applicants like simulations.” Generally, research results show that these statements apply to both high-fidelity and low-fidelity simulations. Future research should focus on comparative evaluations of simulations, the effects of structuring simulations, and the cross-cultural transportability of simulations.

**Key Words:** work samples, assessment centers, situational judgment tests, high-fidelity simulations, low-fidelity simulations

When we organize a research seminar on simulations and search for relevant research articles for graduate students to read, the articles of Wernimont and Campbell (1968), Robinson (1981), and Schmitt and Ostroff (1986) come immediately to mind. Indeed, these articles exemplify the basic idea of behavioral consistency that grounds simulations. Simulations represent more or less exact replicas of tasks and knowledge, skills, and abilities (KSAs) required in actual work behavior. For instance, in the Robinson (1981) study, various simulations for selecting a construction superintendent in a small business setting were carefully constructed. Examples were a blueprint reading task, a “scrambled subcontractor” task, a construction error recognition task, and a scheduling task. Hence, traditional simulations have high stimulus fidelity, response fidelity, and an open-ended nature because they require actual behavioral responses from candidates (Thornton & Rupp, 2006). In this high-fidelity format, simulations have been frequently used for different purposes (e.g., selection, training, job allocation, licensing, credentialing). In Industrial-Organizational (I/O) psychology, simulations are often referred to as “work samples” and “assessment center (AC) exercises,” whereas in educational psychology the terms “performance assessment,” “performance tests,” or “authentic assessment” have been popular (Lane & Stone, 2006). In addition, simulations are also often used as criteria due to their close resemblance to the actual job.

However, in the past decade, this typical face of simulations has changed due to the surge of interest and growing popularity of situational judgment tests (SJTs). Although not a new invention (SJTs already existed prior to World War II), they were reintroduced by Motowidlo et al. (1990) who framed them...
as “low-fidelity simulations.” SJTs score considerably lower on both stimulus and response fidelity than the traditional high-fidelity simulations. They confront applicants with written or video-based descriptions of job-related scenarios and ask them to indicate how they would react by choosing an alternative from a list of predetermined responses (close-ended task, McDaniel, Hartman, Whetzel, & Grubb, 2007; Motowidlo et al., 1990; Weekley, Ployhart, & Holz, 2006).

In this chapter, we review research on the more traditional high-fidelity simulations and contrast it to the growing body of research on low-fidelity simulations. To this end, we structure our review along the following five blanket statements about simulations.

1. “The use of simulations enables organizations to make predictions about a broader array of KSAOs.”
2. “We don’t know what simulations exactly measure.”
3. “When organizations use simulations, the adverse impact of their selection system will be reduced.”
4. “Simulations are less fakable than personality inventories.”
5. “Applicants like simulations.”

Although these statements have been generally accepted as conventional wisdom about traditional high-fidelity simulations, questions can be raised as to whether they are still “no-brainers” with the advent of low-fidelity simulations. Hence, by reviewing research on the full range of simulations including low-fidelity simulations (situational judgment tests) and high-fidelity simulations (assessment center exercises and work samples) our first objective consists of bringing these two streams of research closer together. So far, research on low-fidelity and high-fidelity simulations has followed largely independent paths. As a second objective, we start integrating research on low-fidelity and high-fidelity simulations. In particular, we identify communalities and differences in the two research traditions. At the same time, new integrative research avenues are proposed.

The structure of this chapter is as follows. First we describe the main characteristics of low-fidelity and high-fidelity simulations. Next, the five main questions above constitute the core of our chapter. Finally, we discuss various directions for future research on simulations.

Characteristics of Simulations

Simulations represent contextualized selection procedures that psychologically or physically mimic key aspects of the job. Simulations can be characterized in terms of various features (Callinan & Robertson, 2000; Goldstein, Zedeck, & Schneider, 1993; Whetzel & McDaniel, 2009). Below we discuss the standing of low-fidelity and high-fidelity simulations on these seven features: behavioral consistency, content sampled, fidelity, interactivity, standardization, scoring, and cost/scope.

Behavioral Consistency

Simulations are based on the notion of behavioral consistency (Motowidlo et al., 1990; Thornton & Cleveland, 1990; Wernimont & Campbell, 1968). That is, they are based on the assumption that candidates’ performance on the selection instrument will be consistent with their potential performance on the job. To this end, simulations aim to maximize the point-to-point correspondence with the criterion.

This behavioral consistency logic is conceptualized differently in high-fidelity and low-fidelity simulations. In high-fidelity simulations (AC exercises and work samples), assessors observe and rate actual on-going candidate behavior. This key focus on actual behavior is central to the AC paradigm, as exemplified by the recent AC guidelines (Alon et al., 2009). It is then assumed that the behavior shown by candidates in AC exercises will be consistent and predictive of later job behavior. Conversely, low-fidelity simulations sample applicants’ procedural knowledge about effective and ineffective courses of action in job-related situations such as those described in an SJT (e.g., how to deal with interpersonal situations, decision-making situations, problem-solving situations; Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006a, 2006b). It is then expected that procedural knowledge of effective behavior might be a precursor of showing that effective behavior on the job. Thornton and Rupp (2006) summarized this difference by positing that high-fidelity simulations generate behavioral samples, whereas low-fidelity simulations capture behavioral intentions and knowledge (see also Ryan & Greguras, 1998).

Content Sampled

Simulations can best be conceptualized as multidimensional “methods,” namely methods for measuring a variety of performance dimensions.
(Arthur, Day, McNelly, & Edens, 2003; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel & Whetzel, 2005; Weekley & Jones, 1999). This distinguishes them from more traditional selection tools such as cognitive ability tests or personality inventories. That is also the reason why simulations are sometimes called “method-driven” predictors instead of “construct-driven” predictors.

Thus, the content being sampled in simulations constitutes one major characteristic on which simulations might differ. For instance, in recent years, in the low-fidelity arena SJTs have been developed to capture domains as diverse as teamwork knowledge (McClough & Rogelberg, 2003; Morgeson, Reider, & Campion, 2005; Stevens & Campion, 1999), aviation pilot judgment (Hunter, 2003), team roles (Morgeson, Reider, & Campion, 2005), emotion management (Blickle et al., 2009; MacCann & Roberts, 2008), employee integrity (Becker, 2005), call center performance (Konradt, Hertel, & Joder, 2003), proactivity (Chan & Schmitt, 2000), personal initiative (Bledow & Frese, 2009), goal orientation (Westring et al., 2009), and academic performance (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Peeters & Lievens, 2005). Recently, Christian et al. (2010) developed a taxonomy to categorize these various domains. Their categorization showed that most SJTs capture leadership skills (37.5%), heterogeneous content (33.09%), and interpersonal skills (12.5%). In addition, SJTs were also developed to measure basic personality tendencies (9.56%), teamwork skills (4.4%), and job knowledge and skills (2.94%). This large variety is also striking in high-fidelity simulations such as ACs. Arthur et al. (2003) identified 168 different labels for AC dimensions and their taxonomy classified all of these dimensions into seven broad constructs: consideration/awareness of others, communication, drive, influencing others, organizing and planning, problem solving, and tolerance for stress/uncertainty.

Note that even though both SJTs and ACs might sample the same criterion domains (e.g., performance dimensions such as sensitivity, communication, stress resistance), this does not mean that SJT items and AC exercises measure the same “constructs.” As noted above, SJT items assess whether people know what is the most sensitive, communicative, or stress-resistant option, whereas AC exercises examine whether their actual verbal and nonverbal behavioral manifestations are also sensitive, communicative, or stress resistant.

**Fidelity**

According to I. L. Goldstein et al. (1993) psychological fidelity refers to the degree to which (1) the KSAs required by a job are tapped in the test (regardless of the testing mode), (2) the testing mode is representative of the way tasks are accomplished on the job, and (3) the test captures and scores KSAs not even required by the job. Psychological fidelity differs from physical fidelity, which is defined as the degree to which the simulation simulates actual job tasks.

Basically, the conceptualization of psychological fidelity emphasizes both stimulus (task) fidelity and response mode fidelity. The “fidelity of the task stimulus” refers to the extent to which the format of the tasks and the KSAs required to accomplish the tasks are consistent with how the situation is encountered in the workplace. Simulations might vary in terms of the fidelity with which they present those stimuli. In low-fidelity simulations, the situations might be presented in a paper-and-pencil (written) mode. Accordingly, an SJT takes the form of a written test as the scenarios are presented in a written format and applicants are asked to indicate the appropriate response alternative. Hence, written SJTs have low stimulus fidelity. In video-based or multimedia SJTs, stimulus fidelity is enhanced as a number of video scenarios describing a person handling a critical job-related situation are developed (McHenry & Schmitt, 1994). At a critical “moment of truth,” the scenario freezes and applicants are asked to choose among several courses of action. Thus, video-based and multimedia SJTs allow the item context to be richly portrayed, thereby increasing their stimulus fidelity. Funke & Schuler (1998; Olson-Buchanan & Drasgow, 2006). Recently, organizations have even explored the use of virtual and 3D animated characters in SJTs (Fetzer, Tuzinski, & Freeman, 2010). This recent development comes close to the high stimulus fidelity in AC exercises. In AC exercises, “live” and constantly changing stimuli (confederates, other assesses) typically occur. However, the materials (e.g., in-basket mails) presented might be different from the actual information presented in the real job. In work samples, the level of fidelity might be the highest because candidates are often confronted with the physical stimuli and hands-on tasks that are replicas of the real job tasks (e.g., proofreading and typing in administrative jobs).

Apart from stimulus fidelity, simulations also differ in terms of response fidelity. This component of fidelity denotes the degree to which the response
mode of the candidates is representative of the way they will respond in the actual job. The response fidelity of low-fidelity simulations is typically lower because they have a close-ended (multiple-choice) item format. This means that applicants have to select one response alternative from a list of different response options instead of generating their own solution. The range of response options might vary from one (as in the recently developed single response SJTs of Motowidlo, Crook, Kell, & Naemi, 2009) to more than 10 (as in the tacit knowledge inventories of Wagner & Sternberg, 1985). These response alternatives might also be presented in either a written (lower response fidelity) or a video-based (higher response fidelity) format.

This cued and structured response format feature discriminates low-fidelity simulations from their high-fidelity counterparts such as AC exercises or work sample tests that provide applicants with the opportunity to respond in a manner mimicking actual job behavior. The open-ended format also provides candidates with the discretion to generate their own solutions instead of being constrained to choose one of the predetermined response options. That is also why the most recent AC guidelines do not consider computerized in-baskets (with MC options) to be “true” AC exercises (Alon et al., 2009).

**Interactivity**

Simulations also differ in terms of their degree of interactivity, which refers to the level to which candidate responses are required on the basis of dynamic situational cues instead of on a static situation. Most high-fidelity simulations and especially interpersonally oriented AC exercises such as role-plays, oral presentations, fact-findings, and group discussions are inherently interactive as the candidates have to interact with role-players, resource persons, or other candidates who interfere with the candidates. Accordingly, the next question presented to a candidate might depend on what she or he answered to a prior question. The same is true for interpersonal work samples such as a simulated telephone call. Conversely, individual AC exercises such as in-baskets and planning exercises score lower on interactivity as the materials provided to candidates are static and the presentation of subsequent items is typically not contingent on earlier candidate responses. The same can be said of psychomotoric work samples in which candidates often have to follow specific standardized procedures for accomplishing the job-related tasks.

Traditionally, low-fidelity simulations score low on interactivity as most SJTs are linear. That is, all applicants receive the same set of predetermined item situations and item options. So, the presentation of items is not dependent on their responses to previous items. In some SJTs, however, the applicant’s response to a situation determines the next situation that is presented. So, applicants are confronted with the consequences of their choices. This modality implies that all applicants do not respond to the same items. These SJTs are called “branched,” “nested,” or “interactive” SJTs (Kanning, Grewe, Hollenberg, & Hadouch, 2006; Olson-Buchanan et al., 1998). The technological ability of developing interactive SJTs is possible in multimedia SJTs that present different video fragments to an applicant, based on the applicant’s response to earlier video fragments. This allows the SJT to better simulate the dynamics of interaction. Similarly, computerized versions of individual AC exercises (e.g., PC in-baskets) have tried to incorporate interactivity by presenting information during the simulation (e.g., new incoming mail). However, it should be noted that in most cases this information is not tailored to prior candidate responses (see Lievens, Van Keer, & Volckaert, 2010, for an exception). In the future, this might also become possible with truly adaptive simulations (also known as “serious games”) in which candidates are immersed in a virtual world of work where the computer in real time responds to their actions and automatically scores these actions (Fetzer, 2011).

**Standardization**

Situation (interview question) and response scoring standardization are two key dimensions that distinguish structured from unstructured interviews (Huffcutt & Arthur, 1994). Likewise, the amount of standardization differs across simulations. In low-fidelity simulations, candidates have to select the correct answer from a limited set of predetermined response options. Therefore, in these simulations situation standardization is ensured because all applicants typically receive the same situations and response options (with the exception of branched SJTs, see above). In addition, the same scoring key (i.e., evaluative standard) is determined a priori and used across all applicants, guaranteeing response scoring standardization.

High-fidelity simulations also try to maximize stimulus standardization and scoring standardization. However, their higher level of response fidelity

---

**Interactivity**

Simulations also differ in terms of their degree of interactivity, which refers to the level to which candidate responses are required on the basis of dynamic situational cues instead of on a static situation. Most high-fidelity simulations and especially interpersonally oriented AC exercises such as role-plays, oral presentations, fact-findings, and group discussions are inherently interactive as the candidates have to interact with role-players, resource persons, or other candidates who interfere with the candidates. Accordingly, the next question presented to a candidate might depend on what she or he answered to a prior question. The same is true for interpersonal work samples such as a simulated telephone call. Conversely, individual AC exercises such as in-baskets and planning exercises score lower on interactivity as the materials provided to candidates are static and the presentation of subsequent items is typically not contingent on earlier candidate responses. The same can be said of psychomotoric work samples in which candidates often have to follow specific standardized procedures for accomplishing the job-related tasks.

Traditionally, low-fidelity simulations score low on interactivity as most SJTs are linear. That is, all applicants receive the same set of predetermined item situations and item options. So, the presentation of items is not dependent on their responses to previous items. In some SJTs, however, the applicant’s response to a situation determines the next situation that is presented. So, applicants are confronted with the consequences of their choices. This modality implies that all applicants do not respond to the same items. These SJTs are called “branched,” “nested,” or “interactive” SJTs (Kanning, Grewe, Hollenberg, & Hadouch, 2006; Olson-Buchanan et al., 1998). The technological ability of developing interactive SJTs is possible in multimedia SJTs that present different video fragments to an applicant, based on the applicant’s response to earlier video fragments. This allows the SJT to better simulate the dynamics of interaction. Similarly, computerized versions of individual AC exercises (e.g., PC in-baskets) have tried to incorporate interactivity by presenting information during the simulation (e.g., new incoming mail). However, it should be noted that in most cases this information is not tailored to prior candidate responses (see Lievens, Van Keer, & Volckaert, 2010, for an exception). In the future, this might also become possible with truly adaptive simulations (also known as “serious games”) in which candidates are immersed in a virtual world of work where the computer in real time responds to their actions and automatically scores these actions (Fetzer, 2011).

**Standardization**

Situation (interview question) and response scoring standardization are two key dimensions that distinguish structured from unstructured interviews (Huffcutt & Arthur, 1994). Likewise, the amount of standardization differs across simulations. In low-fidelity simulations, candidates have to select the correct answer from a limited set of predetermined response options. Therefore, in these simulations situation standardization is ensured because all applicants typically receive the same situations and response options (with the exception of branched SJTs, see above). In addition, the same scoring key (i.e., evaluative standard) is determined a priori and used across all applicants, guaranteeing response scoring standardization.

High-fidelity simulations also try to maximize stimulus standardization and scoring standardization. However, their higher level of response fidelity
and interactivity often makes this a challenge. Due to the higher level of response fidelity, candidates can come up with their own (sometimes unusual) responses and solutions. In addition, due to the interactivity AC exercises might often take unforeseen turns so that, for instance, no two group discussions are alike. To counter these potential decreases in standardization and increase the consistency of stimuli that applicants might encounter, current guidelines prescribe that exercise instructions should be carefully constructed and pretested and that role-players should be thoroughly trained (Alon et al., 2009). Additionally, the use of calibrated and trained assessors should ensure that the same evaluative standards are used across applicants. This is especially important in work samples used for certification, credentialing, or licensing purposes.

**Scoring**

Given the contextualized nature of simulations there is often no single correct answer. As the correct response typically depends on the context (organization, job, culture), consensus scoring is often used. In high-fidelity simulations, this is best exemplified by the frame-of-reference training protocol provided to assessors (Lievens, 2001; Schleicher, Day, Mayes, & Riggio, 2002). In such training, considerable effort is undertaken to impose the same evaluative standards (“theory of performance”) on human assessors. As the “live” observation and rating of candidates are cognitively demanding tasks, the AC guidelines further suggest that the carefully selected and trained assessors use various rating aids (e.g., use of behavioral checklists, videotaping of assessee performances). In work samples, the scoring rubric is also often very detailed as raters are expected to score candidates on whether they performed a series of tasks and procedures. In computerized work samples, this detailed scoring logic might be implemented in an automated scoring algorithm.

Many low-fidelity simulations are also based on consensus scoring (Motowidlo & Beier, 2010). However, the implementation of this scoring rubric differs from high-fidelity simulations. In low-fidelity simulations, the scoring key is developed a priori by subject matter experts. There is no “live” rating of behavior and there are no assessors or raters who evaluate the candidates’ on-going behavior. Apart from subject matter experts, SJTs might also be scored on the basis of theoretical and/or empirical grounds (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Olson-Buchanan et al., 1998).

**Cost and Scope**

The costs involved in developing and administering simulations are a final practically important dimension for contrasting low-fidelity and high-fidelity simulations. All simulations are relatively costly to develop as they require the input from multiple groups of subject matter experts to construct the stimulus materials. For instance, AC exercises typically consist of detailed exercise descriptions, role-player guidelines, and assessor checklists (Thornton & Mueller-Hanson, 2004). Similarly, traditional SJTs require subject matter input for constructing the item stems, item responses, and scoring. When video-based SJTs are used, development costs often triple as compared to paper-and-pencil SJTs.

In terms of administration costs, there are marked differences between high-fidelity and low-fidelity simulations. Low-fidelity simulations (especially paper-and-pencil SJTs) enable quick administration across a large number of applicants at once and over the Internet. Along these lines, recent research confirms the equivalence between written and web-based SJTs (Ployhart, Weekley, Holtz, & Kemp, 2003; Potosky & Bobko, 2004). That is also the reason why they are often used as supplements to cognitive ability tests for screening (selecting out applicants in early selection stages). Conversely, the more expensive high-fidelity simulations such as ACs and work samples are typically used in a smaller preselected applicant pool in a specific location. Given these cost and practical advantages, low-fidelity simulations (SJTs) have gained in popularity in large-scale high-stakes selection settings (Lievens, Buyse, & Sackett, 2005a; Oswald et al., 2004; Ployhart & Holtz, 2008).

**Conclusions**

The discussion above shows that there exist marked differences between the traditional simulations with higher fidelity (AC exercises and work samples) and the more recent simulations with lower fidelity on each of the seven features. In addition, our discussion of the features of simulations shows that high-fidelity and low-fidelity simulations should not be regarded as a dichotomy as there exist various intermediate levels of fidelity. All of this raises the question as to how these different characteristics might affect important selection-related outcomes such as criterion-related validity, incremental validity, adverse impact, construct-related validity, fakability/coachability, and applicant perceptions. Therefore, the next sections contrast the
common knowledge on how high-fidelity simulations perform on these critical selection-related outcomes to the growing body of research concerning low-fidelity simulations.

“The Use of Simulations Enables Organizations to Make Predictions about (a broader Array of) KSAOs”

The Criterion-Related Validity of Simulations

Both low-fidelity and high-fidelity simulations are also referred to as “alternative” or “alternate” measures. This label originates from the main reasons behind their emergence. Over the past decade, organizations have relied on simulations to broaden the constructs measured in their selection systems. In particular, they have undertaken substantial efforts to “go beyond” cognitive ability. This endeavor is motivated by at least two reasons, namely the need to reflect the multidimensionality of performance in selection procedures and the need to reduce adverse impact (Goldstein, Zedeck, & Goldstein, 2002; Sackett, Schmitt, Ellingson, & Kabin, 2001).

Against this backdrop, it is not surprising that many studies have examined the criterion-related validity of low-fidelity and high-fidelity simulations. In the AC field, the meta-analysis of Hermelin, Lievens, and Robertson (2007) found a corrected validity of 0.28 for the overall assessment rating (OAR) for predicting supervisory-rated job performance. The use of an OAR in ACs is of practical importance especially for ACs conducted for selection and promotion purposes. Conversely, Arthur et al. (2003) reported a meta-analysis of the validity of final AC dimension ratings. They focused on final dimension ratings instead of on the OAR because the OAR is conceptually an amalgam of evaluations on a variety of dimensions in a diverse set of exercises. In addition, scores on final dimensions are particularly useful for developmental ACs wherein detailed dimension-level feedback is given to candidates. Key results were that several individual dimensions produced validities comparable to the validity of the OAR and the validity of a regression-based composite of individual dimensions (0.45) clearly outperformed the validity of validity estimates of the OAR. Problem solving (0.39), influencing others (0.38), organizing and planning (0.37), and communication (0.33) accounted for most of the variance.

Given that dimension scoring is not common in work samples, meta-analyses on work samples have used overall work sample scores as predictors. Nevertheless, the validity estimate of work samples has also been updated in recent years. Revisiting the validity of work sample tests, two meta-analytic estimates appeared in 1984: an estimate of 0.54 by Hunter and Hunter (1984) and an estimate of 0.32 by Schmitt, Gooding, Noe, and Kirsch (1984), both corrected for criterion unreliability. The 0.54 value has subsequently been offered as evidence that work samples are the most valid predictors of performance yet identified. Roth, Bobko, and McFarland (2005) documented that the Hunter and Hunter estimate is based on a reanalysis of a questionable data source, and report an updated meta-analysis that produces a mean validity of 0.33, highly similar to the prior value of 0.32 of Schmitt et al. Thus the validity evidence for work samples remains positive, but the estimate of their mean validity needs to be revised downward.

Attesting to the increasing interest in low-fidelity simulations, three meta-analyses of the criterion-related validity of SJTs have been conducted in the past decade. McDaniel et al. (2001) conducted the first meta-analysis of the validity of SJTs in employment settings. They reported a mean corrected correlation between SJTs and job performance of 0.34. The second and more recent meta-analysis by McDaniel et al. (2007) found a mean corrected validity of 0.26. Third, similar to the most recent meta-analysis on ACs of Arthur et al. (2003), Christian et al. (2010) conducted a meta-analysis of the validity of SJTs for overall job performance and specific performance domains (e.g., contextual, task, and managerial performance). They found that the corrected validity of SJTs for predicting overall job performance ranged from 0.19 (SJTs measuring job knowledge and skills) to 0.43 (SJTs measuring personality-like tendencies). Results further showed that the validity of SJTs was higher for predicting conceptually related performance dimensions. That is, an SJT measuring interpersonal skills showed higher relationships with contextual performance than an SJT with heterogeneous content, underscoring the importance of predictor–criterion matching.

In sum, research evidence supports the concept that high-fidelity as well as low-fidelity simulations exhibit useful levels of validity. One noteworthy trend across both low-fidelity and high-fidelity simulations is that validities are higher when dimension scores are used as predictors instead of overall scores. In other words, these results exemplify the importance of predictor–criterion matching in
validation of simulations. That is, from a theoretical perspective, these studies support the importance of (1) making fine distinctions within both the predictor and criterion space and (2) aligning specific predictors to specific criteria.

Inspection of the studies included in these various meta-analyses also shows that concurrent designs with incumbents constituted the typical validation design used for both SJTs and work samples. For instance, the number of predictive studies in the SJT meta-analyses was very scarce as only 6 out of 102 studies (McDaniel et al., 2001) and only 4 of the 118 studies (McDaniel et al., 2007) included were predictive validity studies. A similar lack of applicant studies was noted in the work sample domain as the meta-analysis of Roth et al. (2005) retrieved only one applicant study. This lack of validation designs with actual applicants is important as validities seem to differ. In the McDaniel et al. (2001) meta-analysis, the mean validity obtained in predictive studies (0.18) was lower than the validity for concurrent studies (0.35). Clearly, more studies in operational settings are needed to bolster our understanding of the level of predictive validity that might be anticipated in the operational use of simulations. This is especially needed in the SJT and work sample domain. The situation is different in the AC field where the majority of studies have been predictive studies with actual applicants (Arthur et al., 2003).

The use of short time frames is another characteristic of the validation research on SJTs and work samples. In concurrent studies, criterion scores have been typically obtained from both newly selected individuals as well as individuals of varying tenure levels. In addition, in the scarce predictive validation studies the time spans over which criteria have been gathered rarely exceeded a year or two; in most cases they are merely a few months. So, future studies should examine the validities of SJTs and work samples in the long run. In other fields, there has been support for changing validities. For instance, there exists evidence for a declining trend in the validities of cognitive ability (Barrett, Phillips, & Alexander, 1981; Campbell & Knapp, 2001; Deadrick & Madigan, 1990; Schmidt, Outerbridge, Hunter, & Goff, 1988) and an increasing trend in the validities of personality (Lievens, Ones, & Dilchert, 2009). In the AC domain, Jansen and Stoop (2001) discovered that only interpersonal skills in an AC became more valid as individuals progressed through their career (see also Jansen & Vinkenburg, 2006). This longitudinal validation research in the AC field might serve as inspiration for long-term validation research on SJTs and work samples.

The Incremental Validity of Simulations

Although the aforementioned studies in the SJT, AC, and work sample domains show that all of these simulations are valid selection procedures, they do not answer the question as to whether these simulations broaden the constructs being measured. From a utility standpoint, the use of additional predictors is of value only when they explain variance in the criterion beyond that accounted for by other existing predictors. Therefore, another strand of research has examined the incremental validity of each of these simulations over extant measures in the cognitive and personality realm.

Regarding high-fidelity simulations, meta-analysis on work samples of Roth et al. (2005) showed that work samples explained 6% incremental variance over cognitive ability. No estimates for the incremental value of work samples over personality were available. The incremental validity of ACs over and above cognitive ability and personality has received more attention. Although prior meta-analyses using the OAR (Collins et al., 2003; Schmidt & Hunter, 1998) explained additional variance as little as 1%, recent research using overall dimension ratings has painted a brighter picture. Dilchert and Ones (2009) compared the incremental validity of the overall assessment rating versus overall dimension ratings in terms of their incremental validity over cognitive ability and personality in two large managerial samples. Whereas the overall assessment rating did not have incremental validity over cognitive ability and personality, a regression-based composite of AC dimensions offered useful levels of incremental validity (12%). In particular, the dimensions of Influencing Others and Communication provided the largest incremental value when added to personality and ability tests. Recent meta-analytic evidence about the incremental validity of ACs comes from Meriac, Hoffman, Woehr, and Fleisher (2008). They showed that none of the AC dimensions shared more than 12% variance with cognitive ability and personality. Hierarchical regressions on a meta-analytically derived correlation matrix of AC dimensions, personality, cognitive ability, and job performance revealed that adding all AC dimensions provided about 9.7% extra variance above personality and cognitive ability. When controlling for these two traditional tests, organizing and planning
emerged as the single AC dimension with the largest incremental value (5.3%). When only cognitive ability is partialled out, other primary research has shown that the AC method has value, especially for measuring interpersonal, team work, and communication abilities (Dayan, Kasten, & Fox, 2002; Krause, Kersting, Heggstad, & Thornton, 2006).

Regarding low-fidelity simulations, the meta-analysis of McDaniel et al. (2007) revealed that SJTs accounted for 3% to 5% of incremental variance over cognitive ability, 6% to 7% over personality, and 1% to 2% over both cognitive ability and personality. In addition to this meta-analytic evidence in employment settings, there is also substantial evidence that SJTs have value for broadening the type of skills measured in college admission (Lievens et al., 2005a; Lievens & Sackett, in press; Oswald et al., 2004; Schmitt et al., 2009).

In sum, research shows that each type of simulation enables organizations to go beyond traditional ability and personality measures, with the incremental validity of high-fidelity simulations being somewhat higher than that of low-fidelity simulations. Given that in most meta-analyses job performance served as the criterion (see Christian et al., 2010, for an exception), even higher incremental validity results might be obtained when separate criterion dimensions (e.g., contextual performance) are used as criteria. So, the predictor–criterion matching logic might also be fruitfully applied here. Although the incremental validity of simulations over traditional measures seems to be well established, one striking gap is that there are no comparisons of low-fidelity and high-fidelity simulations. Indeed, research on low-fidelity and high-fidelity simulations seems to have been conducted independently. We come back to this issue later.

**The Search for Moderators**

A last group of studies has aimed to further increase the validities of simulations by scrutinizing the effects of potential moderators of the validities obtained. Such studies also advance our understanding of the features of simulations that are responsible for their validity. Regarding low-fidelity simulations in particular, various interesting insights have been gathered in recent years. Most successful moderators being identified pertain to increasing the job relatedness of low-fidelity simulations. One example is that a key moderator of the validity of SJTs emerging in the McDaniel et al. (2001) meta-analysis was whether a job analysis was used to develop the SJT. When SJTs were based on a job analysis, they evidenced higher validities than SJTs that were not based on a job analysis (0.38 versus 0.29). Another example of efforts to increase the fidelity of SJTs consisted of examining whether SJTs with a video-based stimulus format had higher criterion-related validity than those with a paper-and-pencil stimulus format. On the one hand, video-based (multi-media) SJTs might have higher fidelity because the information presented is richer and more detailed, leading to a better match with the criterion behavior as presented on the job. This should result in higher criterion-related validity. However, on the other hand, as cognitive ability is an important predictor of job performance, video-based and multimedia SJTs might be less valid because they are less cognitively loaded (i.e., lower reading component). Furthermore, the video-based format might insert irrelevant contextual information and bring more error into SJTs, resulting in lower validity. Christian et al. (2010) put these competing explanations to the test and conducted a meta-analysis of the criterion-related validities of video-based versus paper-and-pencil SJTs. Results showed that the validities of video-based SJTs of interpersonal skills (0.47) outperformed those of paper-and-pencil SJTs (0.27). A similar trend was observed for SJTs measuring leadership and SJTs with heterogeneous content. One limitation of this meta-analysis was that the content of the SJT was not held constant. That is, the two formats also differed in terms of the content sampled by the SJTs. A primary study of Lievens and Sackett (2006) addressed this confound and discovered that changing an existing video-based SJT to a paper-and-pencil one (keeping content constant) significantly reduced the criterion-related validity of the test. This study also confirmed that scores on the paper-and-pencil presentation of the SJT items had a higher cognitive loading than those on the video-based presentation of the same SJT items.

Apart from moderators involving the fidelity of SJTs, other studies have examined whether the type of SJT response instructions matters in terms of criterion-related validity. The taxonomy of McDaniel et al. (2007) made a distinction between SJTs with a knowledge-based format (“What is the correct answer?”) and SJTs with a behavioral tendency format (“What would you do?”). Similar to the video issue, there are also competing arguments for both the validity of knowledge-based and behavioral tendency instructions. On the one hand, it could be argued that knowledge-based instructions that
invoke maximal performance conditions and have a higher cognitive loading will yield higher validity given the strong record of knowledge tests. On the other hand, it might be posited that behavioral tendency instructions that invoke typical performance conditions and have a higher personality loading will produce higher validity as they better capture behavioral intentions. In line with these arguments, meta-analytic results showed that both formats produced similar validities (0.26). Lievens, Sackett, and Buyse (2009) confirmed these findings in actual high-stakes selection settings. Although knowledge-based instructions had a higher cognitive loading than behavioral tendency instructions, there was no difference between the criterion-related validity of the SJTs under the two response instruction sets.

A final moderator that has been examined pertained to the type of scoring used in SJTs. Bergman et al. (2006) compared 11 different keys (e.g., expert-based, theoretical, hybrid) for scoring an SJT measuring leadership. Results showed considerable variations in the validities obtained, although it was not possible to draw clear conclusions on the superiority of one scoring approach over another one. Recently, Motowidlo and Beier (2010) also highlighted the importance of scoring for the validity of SJTs. They compared three ways of consensus scoring on a managerial SJT: a key based on novices’ judgments, a key based on subject matter experts, and a key based on trait judgments. Results showed that all three keys were valid but that the expert-based key explained incremental variance over and above the other ones.

In sum, in recent years we have seen a large number of studies that have searched for moderators of the validities of low-fidelity simulations. A common thread running through the body of evidence is that increasing the fidelity of low-fidelity simulations (e.g., via use of video items, reliance on job analysis for constructing the items) might increase their validity. Conversely, it is striking that the search for moderators of the validity of ACs seems to have stopped in recent years, even though it should be noted that moderators were already examined in prior meta-analyses (e.g., Gaugler, Rosenthal, Thornton, & Bentson, 1987) and in construct-related AC research (see below). In the work sample domain, Roth et al. (2005) examined a variety of moderators (e.g., type of sample, job complexity) but no moderators emerged. We believe that current moderator research in SJTs might serve as fruitful inspiration for high-fidelity simulations. Similar to SJTs, it might be conceptually useful to break down the high-fidelity simulation method of measurement into stimulus fidelity and response fidelity. Next, we might investigate the separate impact of different sets of exercise instructions and response modes (paper-and-pencil versus oral). Such research might illuminate which specific factors are responsible for the validity of high-fidelity simulations.

“We Don’t Know What Simulations Exactly Measure”

At first sight, this statement might be at odds with the idea underlying simulations. As already noted, simulations aim to be more or less exact replicas of actual job tasks. Accordingly, they are often not originally developed to measure specific constructs. That said, research has tried to ascertain the constructs underlying simulation performance. Two construct-related validation designs have been mostly used. One strand of studies has used an internal validation strategy, whereas another stream of research has adopted an external validation approach.

Internal Validation Research on Simulations

In the internal validation strategy, researchers aim to gather evidence for constructs underlying performance in simulations by investigating whether the structure underlying simulation scores reflects specific constructs. To this end, factor analytic and variance decomposition approaches are typically used. In AC exercises, there exists a voluminous literature that has employed this internal validation approach. Sackett and Dreher (1982) were the first to examine the underlying structure of so-called within-exercise dimension ratings in ACs (i.e., ratings that assessors make on dimensions in each exercise). To this end, they inspected the correlations between ratings on these dimensions across exercises on the one hand and the correlations between ratings within exercises on the other hand. They investigated AC ratings in three organizations. In each of these organizations, they found low correlations among ratings of a single dimension across exercises (i.e., weak convergent validity) and high correlations among ratings of various dimensions within one exercise (i.e., weak discriminant validity). Furthermore, factor analyses indicated that ratings clustered more in terms of exercise factors than in terms of dimension factors. These results were seen as troublesome as they cast doubt on whether AC dimensions are consistently measured across different situations. Many studies...
have replicated the Sackett and Dreher (1982) findings. So far, three studies have been conducted to quantitatively summarize the vast construct-related validity research base (Bowler & Woehr, 2006; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens & Conway, 2001). The most recent review of Bowler and Woehr (2006) used meta-analytical methods to combine 35 AC matrices into one single matrix. Therefore, the confirmatory factor analysis (CFA) results from this meta-analytically derived AC matrix probably represent the best available estimates of exercise and dimension variance. The best fit was obtained for a CFA model with correlated dimensions and exercises. Exercises explained most of the variance (33%), although dimensions also explained a substantial amount of variance (22%). In addition, some dimensions (i.e., communication, influencing others, organizing and planning, and problem solving) explained significantly more variance than others (i.e., consideration of others, drive). It should be noted that dimensions correlated highly (0.79). In another interesting study, Hoeft and Schuler (2001) estimated the amount of variability in AC performance. Their study revealed that AC performance included more situation-specific (57%) than situation-consistent (43%) variability. They also found that candidates performed more consistently on some dimensions than on others. In particular, activity (53%) and oral communication (55%) were the most consistently rated dimensions across exercises.

Over the years, many studies have tried to improve on the measurement of AC dimensions by modifying AC design characteristics (see Lievens, 1998; Lievens & Conway, 2001; Woehr & Arthur, 2003, for overviews). Examples of factors that have been manipulated in past studies include using fewer dimensions, employing behavioral checklists, reducing the assessor– assessee ratio, making dimensions transparent, using psychologists as assessors, providing longer training to assessors, using frame-of-reference training, using task-based

“dimensions,” and providing within-exercise dimension ratings only when assesses have completed all exercises. Although these studies have improved our understanding of which factors affect the quality of construct measurement in ACs, they have typically not produced dramatic changes in the pattern of results. That is, the average correlation between ratings on the same dimension across exercises nearly always stays lower than the correlation of multiple dimensions within a single exercise.

On a broader level, we question whether such dramatic effects might be expected in the first place. This issue of the consistent measurement of AC constructs across different exercises shows marked parallels with the person–situation debate that dominated personality psychology in the late 1960s up through the early 1980s. Today, this debate has settled in favor of interactionism and the average correlation between behavioral observations in one situation and another seems to be only around 0.20 (e.g., Richard, Bond, & Stokes-Zoota, 2003), although the amount of cross-situational consistency has been shown to vary considerably with the characteristics of the situations involved and the behavioral observations being considered. In any case, this amount of cross-situational consistency is similar to the average correlation of an AC dimension across exercises. Moreover, the AC findings seem to extend to all fields wherein different constructs are measured in multiple performance-based exercises. For example, the predominance of situation-specific variance over construct variance has been found in studies about patient-management problems for physicians, bar examinations, hands-on science tasks, etc. (see Lievens, 2009a, for an overview). We also note that this construct-related debate has been nonexistent in work sample research. In that domain, the common practice of working with overall scores instead of dimension scores is not seen as troublesome.

Against these considerations, we agree with the statement of Lance, Foster, Gentry, and Thoresen (2004) that “There may be nothing wrong with assessment center’s construct validity after all” (p. 23, see also Lance, 2008). The typical pattern in the correlations among within-exercise dimension ratings seems to basically reflect what might be expected when individuals perform in different situations (Gibbons & Rupp, 2009; Haaland & Christiansen, 2002; Lance et al., 2000; Lievens, 2002).

Although less popular than in the AC field, the internal validation strategy has also been adopted in the SJT domain. Similar to the findings in the AC domain, results have not been supportive of underlying constructs. Factor analyses of SJT items have revealed a complex factor structure (McDaniel & Whetzel, 2005). When the SJT is developed to measure one specific construct, there is no general factor emerging from SJT scores. In cases where the SJT was designed to measure multiple dimensions, the SJT items do not load on their purported factors. That is also the reason why internal consistency
reliability (instead of test–retest or alternate-form reliability) is not a useful reliability index in SJT research. Again, we might question whether these SJT factor analytic results are really troublesome. After all, it should be noted that traditional SJTs are typically developed to sample criterion domains instead of measuring specific constructs. Additionally, SJT items are construct heterogeneous at the item level (Motowidlo et al., 2009; Whetzel & McDaniel, 2009), that is, the items and response options of SJTs do not reflect one specific dimension (e.g., the same item might have response options indicative of empathy and other ones indicative of problem solving). Future research is needed to examine whether the same complex factor structures are found for so-called construct-driven SJTs (e.g., Bledow & Frese, 2009; Motowidlo et al., 2006a). In those SJTs, item options represent different levels of the same construct and are therefore thought to be more homogeneous, which might result in “cleaner” factor structures.

**External Validation Research on Simulations**

In the external validation strategy, final simulation scores (either SJT composite scores or final dimension ratings) are placed in a nomological network with related constructs (e.g., cognitive ability, knowledge, experience, personality). Accordingly, researchers aim to determine the cognitive and non-cognitive constructs/determinants underlying performance in simulation exercises.

As composite scores are the units of analysis in the external validation approach, we might note a possible contradiction between the internal and external validation strategy. That is, it might be posited that for the results of the external validation strategy to be meaningful there should be evidence that these composite scores assess the dimensions consistently across the various situations (either AC exercises or SJT items). As noted above, there is at best weak evidence for such cross-situational consistency. To solve this potential problem, a large number of detailed scores (either SJT item scores or AC within-exercise dimension ratings) are typically collapsed into a final score so that consistency is improved (via the principle of aggregation).

Regardless of this potential consistency (reliability) issue, the external validation approach has provided much more evidence for the construct-related validity of both high-fidelity and low-fidelity simulations. Regarding high-fidelity simulations, we refer to the recent meta-analyses on both criterion-related validity (Arthur et al., 2003) and incremental validity (Meriac et al., 2008) that found meaningful patterns in the relationship between AC dimensions and job performance, personality, and cognitive ability.

Regarding low-fidelity simulations, research has also focused on identifying the cognitive and non-cognitive determinants underlying SJT composite scores (e.g., Schmitt & Chan, 1998; Weekley & Ployhart, 2005). In the meta-analysis of McDaniel et al. (2001), it was found that SJTs show a significant, moderate correlation ($r = 0.46$) with cognitive ability, even though there was substantial variability around this estimate. Along these lines, the meta-analysis of McDaniel et al. (2007) revealed that the type of response instructions affected the cognitive loading of SJTs. That is, SJTs with knowledge instructions had a higher cognitive loading. Alternatively, SJTs with behavioral tendency instructions had a higher personality loading. As noted above, prior research (Chan & Schmitt, 1997; Lievens & Sackett, 2006) has also identified the degree of stimulus fidelity (written versus video-based) as a key factor in determining the cognitive loading of SJTs. Clearly, we need more research on SJT features that might moderate the construct-related validity of SJT items. Some examples are the type of SMEs (peers, supervisors, customers) used for generating critical incidents/scoring key or the level of specificity of the items.

**Toward an Integrative Theory of Performance on Simulations**

One key conclusion from our review of construct-related research on simulations is that the internal validation strategy has typically not established evidence for a distinct and consistent measurement of constructs in both high-fidelity and low-fidelity simulations. Conversely, the external validation strategy has proven relatively successful in uncovering individual difference determinants underlying performance in high-fidelity as well as low-fidelity simulations. That said, it should be noted that this external validation (nomological network) strategy also has limitations. In particular, modern conceptualizations of validity and validation have criticized nomological network approaches as not being informative about “what a test really measures” (Borsboom, Mellenbergh, & Van Heerden, 2004). We agree that it is to be preferred that a theory...
of performance in the selection procedure is first developed and that this theory guides subsequent validation efforts.

In recent years, some key advancements have been made to build a theory of performance in low-fidelity simulations. In various studies, Motowidlo and colleagues (Motowidlo & Beier, 2010; Motowidlo et al., 2006a) developed a theory of performance on SJTs. The central idea is that SJTs capture procedural knowledge, which is decomposed into general knowledge about effective behavior in situations such as those on the job and specific knowledge about effective behavior in particular job situations. The theory further posits that the general domain knowledge can be represented via implicit trait policies (“implicit beliefs about the relation between the expression of personality traits and their effectiveness in situations”). Motowidlo et al. (2006a) theorize, and then offer evidence, that individual personality shapes individual judgments of the effectiveness of behaviors reflecting high to low levels of the trait in question. In recent extensions of the theory (Motowidlo & Beier, 2010), individual differences in ability and experience were added as determinants of the specific job knowledge and general domain knowledge (i.e., implicit trait policies).

This theory of determinants underlying SJT performance is important for various reasons. It explains that SJTs are valid measures because knowledge of effective behavior (both job-specific and more general) might be a precursor of showing effective behavior on the job and in other situations. At a practical level, it suggests that it may prove possible to make inferences about an individual’s personality from an individual’s judgments of the effectiveness of various behaviors (instead of from their self-reports), which might lead to the development of implicit personality inventories. Such implicit personality inventories might also be less fakable than the traditional explicit (self-report) personality inventories.

In high-fidelity simulations, similar theoretical advancements have been made. Specifically, trait activation theory (Tett & Burnett, 2003) has been suggested as a possible framework for better understanding AC performance. This theory explains behavior based on responses to trait-relevant cues found in situations. In trait activation theory terms, dimensions measured in ACs are no longer seen as stable traits. Instead, they are conceptualized as conditional dispositions (Mischel & Shoda, 1995). This means that stable candidate performances on dimensions can be expected only when the exercises elicit similar trait-relevant situational cues. Exercises are no longer viewed as parallel measures but as triggers of trait-relevant behavior. In addition, trait activation theory provides a theoretical explanation for the variability in candidate performances across different AC exercises. It posits that we should expect only strong convergence among dimension ratings between exercises when the exercises elicit similar trait-relevant situational cues (i.e., are high in trait activation potential for that trait).

The application of trait activation theory provides various new opportunities for AC research and practice. In terms of AC research, the critique that ACs are a-theoretical is long overdue. The trait activation explanation provides the opportunity to develop a stronger theoretical basis for ACs. In terms of AC practice, Lievens, Tett, and Schleicher (2009) delineated how trait activation theory might be used in key AC decisions such as selection of dimensions, design of exercises, observation/rating process, assessor selection, assessor training, and development of feedback reports.

Despite this optimism, we note that so far, these two theoretical frameworks have been developed independently from each other. Future theorizing might try to integrate them as this could open a window of possibilities for more theory-based research on simulations as a whole.

“When Organizations Use Simulations, the Adverse Impact of Their Selection System Will Be Reduced”

As a consequence of the increasing representation of minorities (e.g., blacks, Hispanics, and Asians) in the overall applicant pool, both practitioners and researchers attempt to develop selection instruments that minimize subgroup differences in general and adverse impact in particular. Adverse impact occurs when members of minority applicant pools have significantly less chance to be selected than members of the majority applicant pool. As the adverse impact potential increases when subgroup differences are larger, there is a call for the development and use of selection instruments with less racial and gender differences. Furthermore, the creation of selection tools with equal validity but less adverse impact than cognitive ability tests is considered to be one of the most important goals within personnel selection. These rationales have been key reasons for investing in simulations (apart from their promising criterion-related validity, H. W. Goldstein et al.,
Two major research lines can be identified in this domain. One strand of research aims at determining the degree of subgroup differences (in terms of race and gender) across different simulations. Another body of research attempts to identify moderators of the subgroup differences found. Both are reviewed below, together with avenues for future research.

**Simulations and Subgroup Differences**

Although numerous researchers place simulations on a pedestal claiming that they produce only minimal subgroup differences (e.g., Cascio & Phillips, 1979), others have put this statement to the test by gathering empirical evidence via meta-analyses. An example is the meta-analysis on subgroup differences in AC performance of Dean, Bobko, and Roth (2008). Their meta-analysis consisted of 27 studies performed in applicant as well as incumbent samples that yielded 17, 9, and 18 effect sizes for black–white, Hispanic–white, and male–female comparisons, respectively. In contrast to the traditionally positive image ascribed to simulations when it comes to adverse impact, Dean et al. (2008) found somewhat larger subgroup differences than often assumed. The largest mean difference was observed for black–white comparisons in favor of white test-takers ($d = 0.52$). For minority members other than blacks, the adverse impact potential of ACs was less. Dean et al. (2008) observed a rather small effect size ($d = 0.28$) for Hispanic–white comparisons in favor of whites. For female–male comparisons, a minor gender difference favoring women was observed ($d = -0.19$).

Regarding work samples, a similar meta-analysis dedicated to subgroup differences was conducted by Roth, Bobko, McFarland, and Buster (2008). On the basis of 21 applicant sample studies and 19 incumbent sample studies, they found mean black–white differences of 0.73 $SD$, which is consistent with earlier research on the topic (Bobko, Roth, & Buster, 2005) but again in contrast with the conventional belief that simulations generate no or only minor racial subgroup differences. Until now, few studies were conducted concerning gender differences in work sample performance. An exception is Roth, Buster, and Barnes-Farrell (2010) who recently examined this issue in two different applicant samples. When taking the overall work sample performance into account, a limited female advantage was observed ($d = -0.37$ and $d = -0.34$).

Hence, subgroup differences on high-fidelity simulations, although mostly lower than those on cognitive ability tests, are not negligible. The question is whether these findings are also applicable on the newest variation of those simulations, namely low-fidelity simulations. Whetzel, McDaniel, and Nguyen (2008) conducted a meta-analysis to examine the value of SJTs in reducing subgroup differences. With respect to race, differences in mean SJT scores between subgroups were typically smaller than those reported for various ability tests, including cognitive ability. The difference between whites and minority members was without exception in favor of white participants who scored, respectively, 0.38, 0.24, and 0.29 $SD$ higher than black, Hispanic, and Asian participants. With respect to gender, Whetzel et al. (2008) conclude that women in general outperform men on SJTs, although the female advantage in SJT performance was rather limited ($d = -0.11$).

In short, meta-analytic research shows that adverse impact is reduced through the use of simulations. Yet, recent research evidence is less positive than typically assumed as racial differences are not negligible in high-fidelity as well as low-fidelity simulations, which demonstrates that the potential for adverse impact should not be underestimated when using simulations. Our review also showed that black–white subgroup differences associated with high-fidelity simulations (ACs and work samples) were somewhat larger than those associated with low-fidelity simulations (SJTs).

Although subgroup differences on simulations remain smaller than those typically observed in cognitive ability tests, the impact of unreliability on the estimates of simulations should be taken into account. That is, work samples, AC exercises, and SJTs are typically more unreliable than cognitive ability tests. Therefore, it is important to correct the effect sizes associated with simulations for unreliability.

Another possibility for future research is to compare latent mean differences in terms of gender or race. Anderson, Lievens, Van Dam, and Born (2006) conducted such a construct-driven investigation of gender differences in ACs. Latent mean analyses that corrected for unreliability showed that there was a female advantage on constructs reflecting an
interpersonally oriented leadership style and, in contrast to what was expected based on role congruity theory, on drive and determination. This study demonstrates the necessity to focus on latent mean differences instead of focusing only on observed simulation performance differences. Note that our admonition to correct for unreliability is relevant when it is necessary to make an accurate comparison with cognitive ability tests (for scientific purposes).

When simulations are deployed in the field, it should be clear that operational effect sizes count. Furthermore, as most racial studies focus on black–white differences, little is known about subgroup differences related to other subgroups. Thus, future research is needed to determine to what extent the research findings on black–white comparisons can be extended to other racial subgroups (e.g., Arabs, Asians, Indians).

**Moderators of Subgroup Differences**

Apart from examining the size of the subgroup differences, an equally important question concerns the identification of the driving forces between these differences. In recent years, researchers have begun to shed light on possible moderators of racial and gender differences in simulation performance.

Concerning ACs, systematic research concerning the possible influencing factors of the racial and gender performance differences observed has been scarce until now. Exceptions are the studies conducted by Goldstein, Braverman, and Chung (1993) and Goldstein, Yusko, Braverman, Smith, and Chung (1998). Both studies addressed the range in black–white differences across AC exercises instead of focusing solely on the OAR. Although both revealed a varying degree of black–white differences on the different types of AC exercises, Goldstein et al. (1998) made a first attempt to examine why different AC exercises result in different racial subgroup differences. Their study contained 366 employees who participated in an AC consisting of 7 different exercises. Results showed that the size of the black–white performance differences was contingent upon the cognitive loading of the AC exercise. When controlled for cognitive ability, none of the exercises displayed subgroup differences. The role-play, which was the only AC exercise that demonstrated no significant correlation with cognitive ability, was also the only exercise that generated no meaningful black–white differences. Furthermore, Dean et al. (2008) reported that the size of the racial difference in AC performance was moderated by sample type such that larger black–white differences were observed in applicant samples than in incumbent samples. This might suggest that subgroup difference sizes are systematically underestimated in many studies within this domain, as for convenience purposes most of these studies rely on incumbent samples instead of applicant samples.

Research findings from studies aiming to clarify the influencing factors of subgroup differences in work sample performance show numerous parallels with the above-mentioned findings in the AC domain. The first moderator of black–white differences that was identified by Roth et al. (2008) was the sample type. Like the black–white differences in AC performance, subgroup differences in work sample performance proved to be smaller for incumbents than for applicants (d = 0.53 versus d = 0.73). In addition, exercise type and cognitive saturation also played a moderating role. In terms of exercise level, technical and in-basket exercises exhibited larger subgroup differences (d = 0.70) than oral exercises and role-plays (d = 0.20). Furthermore, cognitive test loading proved again to be one of the most important moderators, as larger black–white differences were observed in exercises that measured cognitive and job knowledge skills (d = 0.80) compared to exercises that mainly focused on social skills (from d = 0.21 to d = 0.27).

The small gender difference on high-fidelity simulations in favor of women is suggested to be function of the interpersonal nature of the constructs measured. That is, the female advantage on AC performance might be attributed to their higher levels of affiliation and sensitivity in comparison to men (Roth et al., 2010; Sackett & Wilk, 1994). Furthermore, when gender differences in work samples were analyzed more thoroughly on the exercise level, men usually scored slightly higher on technical exercises, whereas women had a significant advantage on exercises measuring social skills or writing skills.

In the context of the objective of this chapter it is important to examine whether these above-mentioned moderators of subgroup performance differences on high-fidelity simulations generalize to low-fidelity simulations. Although high-fidelity and low-fidelity simulations have different characteristics, we can infer some similar influencing principles for both types of simulations concerning subgroup difference size. In particular, as is the case with high-fidelity simulations, the cognitive loading of an SJT appears to be one of the strongest moderators of the observed subgroup differences in SJT performance.
The meta-analysis of Whetzel et al. (2008) showed that the correlation of SJTs with cognitive ability explained almost all of the variance in mean racial differences across studies. This correlation is particularly strong for black–white differences but can also be observed in other racial subgroup comparisons albeit to a lesser degree. Apart from cognitive ability, other moderators unique to the SJT domain were identified. On the basis of their meta-analysis, Whetzel et al. (2008) suggested the personality loading of the SJT is a second important moderator of subgroup differences. In particular, the observed black–white and Asian–white differences were larger when the emotional stability loading of the SJT was lower. When the SJT had lower conscientiousness and agreeableness loadings, larger Hispanic–white subgroup differences were observed. A third possible moderator of mean race differences identified by Whetzel et al. (2008) was the type of response instruction. Results showed that SJTs with knowledge instructions yielded slightly larger mean race differences than SJTs with behavioral instructions, which is related to the higher cognitive loading of knowledge-based response instructions.

Other primary studies have tried to determine alternative moderators that might influence the adverse impact of low-fidelity simulations. As the presentation mode of SJTs can take many forms, one factor that has emerged as particularly powerful is the format (written versus video-based) of the SJT. Chan and Schmitt (1997) compared a video-based SJT with a paper-and-pencil SJT while keeping the test content of the SJT constant. They found that the video-based SJT generated significantly less black–white differences compared to its paper-and-pencil counterpart. Furthermore, they showed that this observed race × format interaction was attributable to differences in reading comprehension ability and differences in face validity perceptions.

The last search for moderators concerns the subtle gender bias that is observed in SJT performance. Whetzel et al. (2008) suggested that this gender performance difference—as is the case in high-fidelity simulations—might be due to gender differences in terms of the personality traits that are implicitly triggered by the SJT situations as these scenarios are often interpersonal in nature. Indeed, the correlation with personality was found to influence the size of the gender differences so that the higher the conscientiousness loading and the agreeableness loading of the SJT, the better women performed on the test and, consequently, the higher the gender differences (Whetzel et al., 2008). In contrast to racial subgroup differences, gender differences in SJT performance were not significantly moderated by cognitive ability.

Across our review of low-fidelity and high-fidelity simulations, it was striking that similar factors emerged as potential determinants of adverse impact. Specifically, the cognitive test loading (i.e., the nature of the constructs measured) proved to be one of the key moderators of the racial subgroup difference size on simulation performance. The cognitive load theory (Sweller, 1988, 1989) might serve as the theoretical basis for this finding. According to Sweller, tasks vary in the load they impose on the cognitive system (p. 361, Goldstein et al., 1998). Specific task factors (e.g., novelty, structure, time pressure) might influence the cognitive load of an exercise. Future research is necessary to shed light on which explicit simulation factors contribute to the cognitive test loading and on its correlation with subgroup differences. In addition to cognitive test load, the stimulus format in SJTs plays a role.

Although these findings on moderating variables are valuable, they are also somewhat meager. Therefore, we need to continue to accelerate our investigation of potential determinants of adverse impact. One potential unexplored factor might be the response format (constructed versus close-ended, multiple choice). Along these lines, Edwards and Arthur (2007) took the first step toward a thorough examination of the response format as a determinant of subgroup differences, albeit in a knowledge test instead of a simulation exercise. They demonstrated that knowledge tests with an open-ended response format generated significantly less black–white performance differences than its multiple-choice counterpart. Future research should examine whether this research finding is also applicable to SJTs. Furthermore, it is necessary to examine to what extent variations in the response format of high-fidelity simulations result in variations of subgroup performance differences.

“Simulations Are Less Fakable and Coachable Than Personality Inventories”

In simulations, it has generally been assumed that faking is not really an issue. Although the motivation to fake might be as high in simulations as in personality inventories, candidates often do not have the ability to fake due to the cognitive demands of the exercises or due to their own limited proficiency level and behavioral repertoire. Indeed, it might be
more difficult to fake good in AC exercises because of their higher levels of response fidelity and interactivity. That is, candidates are required to show actual behavior and to react directly to unpredictable and live stimuli (e.g., fellow candidates, confederates).

However, with the advent of low-fidelity simulations (SJTs) in the past decade, it seems that the statement that simulations in general are less prone to faking should be toned down because SJTs use a self-report response format. Therefore, faking, practice, and coaching effects have been more frequently investigated for these simulations than for their high-fidelity counterparts. Below we review the available research studies side by side and discuss how research results can be used to cross-fertilize both fields.

Effects of Faking
As it has been assumed that high-fidelity simulations measure maximal performance instead of typical performance (Ployhart, Lim, & Chan, 2001), faking has not been a prevalent research topic in high-fidelity simulations. This does not mean that high-fidelity simulations such as AC exercises might not create demands that invoke behavior among candidates that is not typical of their usual job behavior. For instance, in various studies Kleinmann and colleagues discovered that candidates try to discover the dimensions to be assessed in AC exercises. However, they are generally not very successful at this (see Kleinmann et al., 2011, for an overview). This body of research further showed that the ability to identify the dimensions to be assessed in AC exercises was positively associated with both cognitive ability and performance in the AC exercise. There is also evidence that attempts to fake and manage impressions depend on the type of AC exercises. McFarland, Yun, Harold, Viera, and Moore (2005) studied impression management across various AC exercises and related impression management tactics to the situational demands of the AC exercises. Assessment center exercises that tapped interpersonal skills were less prone to impression management than structured interviews. Among AC exercises, there were also differences, with interpersonal exercises being more prone to impression management effects than technical exercises. Clearly, further research is needed to investigate how candidates approach high-fidelity simulations and which tactics they use to manage the impressions of assessors. Similar to SJTs, more research is needed on AC features that seem to moderate impression management effects.

Contrary to ACs, there is a growing research base regarding faking effects on low-fidelity simulations. Hooper, Cullen, and Sackett (2006) summarized the available research evidence and discovered that differences in mean scores between respondents who were asked to respond as honestly as possible and respondents who were asked to “fake good” varied between 0.08 and 0.89 standard deviations. They also concluded that the SJT faking good effects are considerably smaller than in the case of personality measures. Kanning and Kuhne (2006) drew similar conclusions when comparing response formats of personality inventories and SJTs.

Interestingly, Hooper et al. (2006) also identified several moderators that might make an SJT more fakable and that might explain the large differences across faking studies. First, when SJT items had a stronger cognitive loading, they were less fakable. Second, more transparent items were more fakable. Third, the type of response instructions was a key factor as it affected the cognitive loading and amount of response distortion in SJTs (Nguyen, Biderman, & McDaniel, 2005; Ployhart & Ehrhart, 2003). Behavioral tendency instructions exhibited higher faking than knowledge-based instructions. Finally, the type of study design used played a role. Laboratory findings were a worst-case scenario in comparison to faking in real-life selection. Such experimental laboratory designs manipulate faking and investigate whether applicants can fake a test. This is not the same issue as whether applicants do fake a test in actual selection situations.

So far, the majority of faking studies in the SJT domain have compared mean scores across various groups (faking versus honest; applicants versus incumbents). The effects on criterion-related validity have been largely ignored. One laboratory study showed that faking reduced criterion-related validity from \( r = 0.33 \) to \( r = 0.09 \) (Peeters & Lievens, 2005). Along these lines, the finding of lower mean validity in the small number of existing predictive studies (i.e., mean \( r = 0.18 \) versus 0.35 for concurrent studies) in the McDaniel et al. (2001) review also deserves attention. Although this difference suggests that faking associated with an actual selection environment does not negate the validity of the SJTs, it should be clear that more studies are needed that scrutinize the effects of faking on the criterion-related validity of SJTs.

Research has also started to examine approaches for countering these faking effects in SJTs. One of the most frequently heard suggestions consists of using
knowledge-based instructions as they invoke maximal performance conditions (Lievens, Sackett et al., 2009; Whetzel & McDaniel, 2009). As another strategy for reducing response distortion in SJTs, Lievens and Peeters (2008) hypothesized that asking people to elaborate on their chosen SJT response would reduce faking. This hypothesis was based on social-psychological research on accountability (reason giving). Results showed that response distortion was reduced only when people were knowledgeable and familiar with the situations portrayed in the SJT items. Future studies should investigate the viability of other approaches. For instance, Bledow and Frese (2009) carefully constructed their response options in their SJT about personal initiative. That is, response alternatives that were low on personal initiative were considered to be high on emotional stability, increasing their endorsement frequencies as distractors. It might be worthwhile to examine the effects of faking on such carefully constructed response options. Another possibility consists of comparing the effects of intentional response distortion on single response option SJTs (Motowidlo et al., 2009) to multiple response option SJTs.

**Effects of Retesting**

Apart from response distortion, practice or retest effects have also been examined. Practice effects refer to candidates’ learning from their own experience by taking an alternate form of a test under standardized conditions (i.e., there is no external intervention). Most organizations in the private and public sector have installed retesting policies in promotion and hiring situations. The opportunity for retesting is also mandated for tests used in making admission, licensing, or certification decisions.

The key issue here is whether candidates can improve their scores when they retake the simulations. Contrary to the voluminous literature on retest effects on cognitive ability tests (Hausknecht, Halpert, Di Paolo, & Gerrard, 2007), research on retest effects is scarce in both the low-fidelity and high-fidelity domain. In the AC domain, Kelbetz and Schuler (2002) reported that prior assessment center experience explained 3% of the variance of the OAR. Generally, repeated participation in an AC provided candidates with a gain equivalent to an effect size of 0.40. Regarding SJTs, Lievens, Buyse, and Sackett (2005b) demonstrated that retest effects of SJTs (0.40) were no larger than those of more traditional tests such as cognitive ability or knowledge tests. There were also no differences in terms of the validity of scores of one-time test-takers and repeated test-takers.

To further advance our understanding of retest effects in this domain we recommend that future research go beyond comparisons among mean scores (initial test versus retest). A key problem with the primary reliance on mean score changes is that the reasons why people scored better the second time (e.g., learning of tricks gimmicks, increased test sophistication, or genuine improvement on the construct of interest during the interval between two administrations) remain largely unexplored. Along these lines, Lievens et al. (2005b) developed a framework for disentangling these rival explanations by studying the effects of retesting on both mean score and validity change across a variety of tests. Future studies can adopt this framework to study retesting in both low-fidelity and high-fidelity simulations.

Finally, retesting creates the need for periodic administration of comparable tests. In the measurement field, there exist well-developed technologies for constructing alternate versions of traditional tests (Nunnally & Bernstein, 1994). Central features of these technologies include the development of item pools, pretesting items to gather item statistics, and test construction approaches, which impose various item parameter and content constraints (e.g., specifying the use of items matched on difficulty and discrimination parameters). However, these classic technologies are difficult to apply to simulations because they are multidimensional and typically reflect a construct domain that is not fully understood (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998). As noted above, in simulations, individual items and exercises are commonly designed to sample key job domain aspects, rather than to reflect a clearly understood construct. These concerns have been voiced for both low-fidelity and high-fidelity simulations (Brummel, Rupp, & Spain, 2009; Clause et al., 1998).

To circumvent these problems, various approaches for constructing alternate forms of low-fidelity and high-fidelity simulations have been developed in recent years. Some approaches (Oswald, Friede, Schmitt, Kim, & Ramsay, 2005) use a randomization approach in which a myriad of different SJTs, which satisfy particular specifications, are constructed. Next, SJTs are randomly assigned to participants. Other approaches build on item generation theory (Kyllonen, 2002; Lievens & Sackett, 2007). The hallmark of item generation
theory is that it is possible to determine a priori the factors that contribute to item difficulty. Hereby, the radicals-incidentals approach is often used. Radicals refer to structural item features that determine item difficulty (Irvine, Dann, & Anderson, 1990). Conversely, incidentals refer to changes in surface characteristics of items that do not determine item difficulty. For instance, Clause et al. (1998) successfully applied a cloning (item isomorphic) approach for constructing alternate SJT forms. This meant that alternate SJT items differed only on superficial and cosmetic features (linguistic and grammar changes to item stems and item options). The content and context of the item stems and options were identical.

In high-fidelity simulations, similar cloning approaches have been evaluated. Both Brummel et al. (2009) and Lievens and Anseel (2007) successfully used an incident isomorphic cloning approach for constructing alternate AC exercises. In this approach, the alternate forms do not only differ in terms of cosmetic changes. In addition, the concrete context wherein the critical incidents are embedded (the item stems) and the ways of responding to them (the item options) differ across forms. Only one study has compared the performance of various alternate SJT form development approaches (randomization, incident isomorphic, and item isomorphic) that differed in terms of the similarity of the items included in the alternate SJT forms (Lievens & Sackett, 2007). The approach that built in the least similarity among alternate SJT forms (i.e., random assignment of SJT items across forms) resulted in the smallest retest effects. Conversely, the approach that built in the most similarity across the forms (item isomorphic) resulted in the largest correlation among forms. The various alternate SJT development strategies did not lead to differences in criterion-related validity.

**Effects of Coaching**

Large-scale high-stakes testing situations (e.g., public sector selection, admissions testing) are typically much more open to public scrutiny than employment testing. In those settings, those considering higher education all know well in advance that they will be asked to take a particular test as part of the application process, and a combination of this public knowledge and relatively high testing volumes makes commercial coaching viable. In contrast, job applicants may encounter an enormous array of differing tests as they apply for various jobs, thus limiting the viability of a coaching enterprise in many settings. As simulations are often used in large-scale high-stakes testing situations, knowledge about the effects of coaching on simulations is important. Contrary to retest effects, coaching effects relate to learning through instruction (in the form of an external intervention such as feedback from others, information sharing, tutoring, and test preparation).

Again, research is scarce. Lievens (2001) summarized the literature on the effects of coaching on AC exercises. He found coaching to be more effective (in terms of mean score change) for performance in individual exercises such as in-baskets than for performance in group discussions. However, it is difficult to draw firm conclusions. The coaching literature in AC exercises suffers from both conceptual and methodological problems. Conceptually, a problem with interpreting the results of these coaching studies is that an amalgam of various coaching tactics (e.g., self-study, explanation of constructs measured, modeling, feedback) was studied. Methodologically, it is difficult to draw conclusions on the effects of coaching in field settings as there is no random assignment of participants to the coached and uncoached groups.

With respect to SJTs, Cullen, Sackett, and Lievens (2006) examined the coachability of SJTs developed for consideration as selection instruments in high-stakes testing (college admission process). Strategies for raising scores on each test were generated, and undergraduates were trained in the use of the strategies using a video-based training program. Results indicated that some SJTs were susceptible to coaching. The coaching intervention had no detrimental effect on the validity of the SJTs.

"**Applicants Like Simulations**"

In general, applicants seem to prefer selection tools, which they perceive as job related. That is one of the reasons why simulations typically receive favorable ratings. Research evidence shows that this is especially true for ACs (e.g., Dodd, 1977) and work samples (e.g., Callinan & Robertson, 2000). Consequently, applicants often prefer high-fidelity simulations over other selection tools such as personality questionnaires, biodata, or cognitive ability tests (Hausknecht, Day, & Thomas, 2004; Iles & Mabey, 1993; Rynes & Connerly, 1993).

Although shedding light on the overall applicant perceptions toward selection tools is important, it is equally critical to identify which separate factors
influence these applicant perceptions during the selection procedure. Some researchers made strides forward in clarifying possible factors that might influence applicant perceptions. Hausknecht et al. (2004) reviewed the literature on applicant perceptions and conducted a meta-analysis based on 86 samples. Based on their analysis, job relatedness, face validity, and perceived predictive validity emerged as the most important factors influencing applicant perceptions (Hausknecht et al., 2004). Steiner and Gilliland (1996) also showed that perceived face validity was by far the strongest correlate of favorability perceptions among American and French respondents, which in turn might explain why both samples rated work samples as one of the most favorable selection tools. Smither, Reilly, Millsap, Pearlman, and Stoffey (1993) studied applicant reactions to various selection tools. They found that respondents perceived work samples to be significantly more job related than personality questionnaires, abstract reasoning tests, and biodata. Macan, Avedon, Paese, and Smith (1994) also demonstrated that applicants perceived ACs to be significantly more face valid than cognitive ability tests.

Other researchers proposed that opportunity to perform should be regarded as another important determinant of applicant fairness perceptions, besides job relatedness and face validity (Schleicher, Venkataramani, Morgeson, & Campion, 2006). Opportunity to perform can be defined as the perception that someone had an adequate opportunity to demonstrate his or her knowledge, skills, and abilities in the testing situation (p. 560, Schleicher et al., 2006). Schleicher et al. confirmed the influence of opportunity to perform on procedural fairness perceptions. Results also suggested that test-takers report more opportunity to perform on nonwritten testing tools. Unfortunately, the selection battery studied in Schleicher et al. contained no simulation exercises. Although a large opportunity to perform ratings for simulations can be suspected, future research is necessary to confirm this assumption.

Another possible factor that might determine whether high-fidelity simulations are favorably viewed by applicants is suggested by Potosky (2008). In her conceptual paper on the assessment process she considers the personnel selection process to be a two-way-interaction between the applicant and the organization, in which the test administration medium plays an important role. Potosky (2008) posited that administration media may vary in terms of four distinct attributes: transparency (the degree to which the selection medium facilitates the communication process), social bandwidth (the number of social cues that can be presented by the selection medium), interactivity (the pace of reciprocal exchange between communication parties during the selection), and surveillance (real and/or perceived privacy of the test medium). Selection tools that include face-to-face interactions are expected to score high on all attributes, particularly on social bandwidth and interactivity. As high-fidelity simulations are characterized by their highly interactive nature, this may serve as an explanation for the observed positive applicant perceptions for ACs. However, as Potosky emphasizes, future research should determine the exact standing of different selection tools (e.g., simulations) on these attributes and the effect of varying attribute combinations on applicant perceptions.

Taken together, earlier research proposed job relatedness, opportunity to perform, and the characteristics of the test administration medium as important moderators of the perceived positive applicant perceptions of high-fidelity simulations. As low-fidelity simulations do not include face-to-face interactions, their face validity, opportunity to perform, social bandwidth, and interactivity might be lower than those of ACs and work samples. Therefore, we can question whether applicant perceptions on SJTs will be as positive as those on high-fidelity simulations. Although more studies (and especially comparative studies) on this matter are available, research findings have suggested that a clear distinction in applicant perceptions can be made on the basis of the SJT format used. Along these lines, Chan and Schmitt conducted a pioneering study (1997) in which they compared a video-based SJT and a written SJT. Consistent with earlier research that compared simulations with paper-and-pencil tests (e.g., Smither et al., 1993), results showed that the video-based SJT received significantly higher face validity perceptions than its written counterpart. As Chan and Schmitt (1997)—unlike many previous researchers within this domain—kept the test content constant in both SJT variants, the difference in face validity could be attributable solely to the format change (video-based format versus written format). Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) also investigated the effects of test presentation modality on test-taker perceptions. With test content kept constant, test-taker
perceptions on a paper-and-pencil test, a computerized test (identical to the written test variant, but presented on computer screens with automatic page turner), and a multimedia test (using interactive video material) were compared. Results revealed a clear preference for the multimedia format, which test-takers considered as more face valid, content valid, and more enjoyable. Kanning et al. (2006) drew similar conclusions. They examined the factors of SJT presentation on test-taker perceptions. They concluded that SJTs that are interactive and used a video-based modality for the presentation of stimuli as well as for the response options received the highest ratings as compared to other SJTs that varied in other ways on these factors.

In terms of future research, these findings should encourage researchers to experiment with varying stimulus and response fidelity modes of SJTs (e.g., webcam SJTs) to increase the job relatedness, opportunity to perform, and interactivity of low-fidelity simulations. Another example of such hybrid simulations is the development of open-ended SJTs. Edwards and Arthur (2007) already compared open-ended tests with multiple-choice variations of the same selection tool and demonstrated that the open-ended version generated significantly more positive test-taker reactions.

Another avenue for future research concerns systematic comparisons of applicant reactions to low-fidelity and high-fidelity simulations, which should increase our knowledge about possible moderators of applicant perceptions on simulations. To make useful inferences about possible influencing factors, we should aim to keep the simulation type as constant as possible (Arthur & Villado, 2008). So, comparing test-taker perceptions of a paper-and-pencil in-basket exercise to perceptions of an interactive role-play does not bring much added value to the search for moderators in the applicant perception domain as they mix content and method. Along similar lines, Ryan and Huth (2008) argued that it might be fruitful to use selection attributes as the unit of analysis instead of the selection instrument as a whole. They identified several possible factors of selection tools that might influence test-taker perceptions, while distinguishing between content, format, and context factors.

A last direction for future studies concerns the examination of how test-takers experience completing simulation exercises. Until now we do not know which cognitions and emotions occur during the test-taking. Think-aloud studies might shed some light on how simulation exercises are experienced. Although this research method has already been conducted in the personality and ability domain, its use in the context of different simulations is nonexistent to date (Ployhart, 2006).

**Directions for Future Research**

Throughout this chapter we have mentioned possible areas in need of future research. In this concluding section, we recapitulate the key avenues of dire need of research. Logically, we focus on some common threads running through the previous sections. In particular, we discuss the following domains: (1) comparative evaluations of simulations, (2) structure and simulations, and (3) cross-cultural transportability of simulations.

**Comparative Evaluations of Simulations**

One common thread running through this chapter is that studies that compared low-fidelity and high-fidelity simulations are extremely scarce. So, it is not an exaggeration to posit that research on these different types of simulations has evolved rather independently from each other. Probably, this lack of integrative research on simulations results from the fact that low-fidelity simulations are often considered when cost and time constraints make high-fidelity simulations impractical to develop (Motowidlo et al., 1990). Clearly, future studies are needed to bridge and integrate these two domains.

First, we believe it is important to compare the criterion-related validity of low-fidelity and high-fidelity simulations. Although research has revealed that the reduced fidelity of low-fidelity simulations does not jeopardize their validity and incremental validity over established measures in the ability and personality domain (McDaniel et al., 2007; McDaniel et al., 2001), investigations of whether high-fidelity simulations (AC exercises) have incremental validity over and above low-fidelity simulations (SJTs) are sorely lacking. Although meta-analytic evidence shows both types of simulations are valid, the possible gains in validity that organizations might obtain if they supplement a low-fidelity approach (in early selection stages) with a high-fidelity approach (in later stages) are unknown. Equally important from a utility perspective, we do not know whether these potential gains in validity depend on the criterion constructs (performance dimensions) targeted and whether they remain over time (i.e., when dynamic criteria are used to test the effectiveness of low-fidelity and high-fidelity simulations over
time). From a theoretical perspective, it might also be intriguing to know how procedural knowledge about which behavior to show in a written situation (as captured by an SJT) translates to actual behavior in simulated job-related situations (as captured by work samples and AC exercises) and how both of these predictions relate to supervisory-rated behavior (as captured by job performance ratings).

Apart from criterion-related validity, it is equally important to compare low-fidelity and high-fidelity simulations in terms of other key outcomes such as adverse impact and applicant perceptions. Hereby, it is pivotal that we do not compare an SJT about leadership to AC exercises targeting decision-making skills. Research should try to conduct comparisons of two samples (low-fidelity versus high-fidelity) of the same dimensional space (Arthur & Villado, 2008). For instance, construct-driven comparisons might be possible with construct-oriented SJTs (e.g., Bledow & Frese, 2009; Motowidlo et al., 2006a). These SJTs are specifically designed to include multiple items for specific constructs. Similarly, in the AC domain, there have been calls to use a larger set of shorter exercises that measures only one or two dimensions (e.g., Brannick, 2008; Dilchert & Ones, 2009; Lievens, 2008). If both SJTs and AC exercises are found to measure specific constructs, then a fine-grained comparison of low-fidelity and high-fidelity simulations on the level of dimensional scores is possible (Lievens & Patterson, 2011).

It is important that comparative evaluations of simulations try to go beyond simply demonstrating that “one simulation is better than the other one.” That is, this comparative research should also reveal which specific stimulus, response, and scoring features of simulations (instead of the simulations in general) impact key selection outcomes such as validity, adverse impact, fakability, and applicant perceptions.

**Structure and Simulations**

As noted above, a key difference between low-fidelity simulations such as SJTs and high-fidelity simulations is that low-fidelity simulations have a more structured format. This higher level of standardization is exemplified by the fact that in most low-fidelity simulations all candidates receive the same situations and response options. In addition, a predetermined scoring key is used. Hence, human raters are present only behind the scenes (e.g., as experts deciding on the scoring key). In high-fidelity simulations, an open-ended response format is typically used, with assessors evaluating behavioral responses.

Given that in high-fidelity simulations assessors have to evaluate a rapid stream of behavior much effort has been undertaken over the years in creating a more structured evaluation environment for assessors. Examples are the use of behavioral checklists and training programs. Clearly, this research attention is needed. However, there is also another side of the equation. We might also increase the efforts in better structuring the situations given to candidates in AC exercises.

One approach entails planting structured situational stimuli in AC exercises. In this respect, Brannick (2008) cogently argued “to deliberately introduce multiple dimension-relevant items or problems within the exercise and to score such items” (p. 132). This approach is based on trait activation theory (Lievens, Tett et al., 2009; Tett & Burnett, 2003). Situation relevance is a key concept here. A situation is considered relevant to a trait if it provides cues for the expression of trait-relevant behavior. So, if organizations want to assess candidates on a dimension such as resistance to stress that is related to the trait of emotional stability, they must use exercises that put people in a situation that might activate behavior relevant to the trait of interest (without rating this trait). Let us take an oral presentation with challenging questions as an example. Examples of stimuli to elicit behavior relevant to a dimension such as resistance to stress (a facet of the broader trait of emotional stability) might be the inclusion of a stringent time limit, sudden obstacles, or information overload. When interpersonal exercises (e.g., role-plays and oral presentations) are used, role-player cues are another means for structuring the exercise and eliciting job-related behavior (Schollaert & Lievens, 2011). For example, to arouse behavior related to interpersonal sensitivity, role-players might state that they feel bad about a candidate’s decision. It is important that these role-player cues should subtly elicit asssesee behavior because the situations might otherwise become too strong.

Another approach in the context of creating a more structured test environment in AC exercises consists of increasing the number of situations in ACs. In low-fidelity simulations, candidates typically complete over 40 SJT items. For instance, we might consider including a large number of shorter exercises (exercise “vignettes”) in the AC. Along these lines, Brannick (2008) recommends using five
6-minute role-plays instead of a single 30-minute role-play (e.g., with a problem subordinate) so that samples of performance are obtained on a large number of independent tasks each of which is exclusively designed to elicit behavior related to a specific trait (see also Motowidlo et al., 2006a, for the use of 1-minute or 2-minute role-plays). As another example, we could aim to measure communication by including “speed” role-plays with a boss, peers, colleagues, customers, and subordinates.

Stimuli might also be presented via videotape (Gowing, Morris, Adler, & Gold, 2008). In Lievens (2009b), candidates for police officer jobs watched video-based scenes. Each of these scenes triggered a specific dimension. At the end of each scene, the character in the video spoke directly into the camera. Candidates were next required to answer the character directly, with their verbal and nonverbal reply being captured by a webcam. These reactions were then coded by trained assessors. One set of analyses examined the consistency of assessors’ dimensional ratings across scenes (i.e., convergent validity). That is, did scenes that were developed to trigger a similar dimension provide a consistent measurement of that specific dimension? In line with the expectations, the consistency in assessor ratings was acceptable (only ratings on a more ambiguous dimension such as integrity were slightly less consistent), confirming that the use of multiple videotaped dimension for measuring one dimension might serve as a good vehicle for obtaining a more consistent measurement of the dimension of interest. Brink et al. (2008) also showed candidates short video scenes and asked them to react to what they saw. They focused on the discriminant validity of assessor ratings and found that assessors were able to make better differentiations among the various dimensions. Future applications may enable creation of so-called “serious games” or avatar-based simulation exercises wherein participants take on a virtual identity and are confronted with standardized stimuli in a virtual workplace (Fetzer et al., 2010; Rupp, Gibbons, & Snyder, 2008). Future studies might examine the effects of increasing the structure and behavior elicitation in simulations on reliability and validity.

**Cross-Cultural Transportability of Simulations**

Due to the globalization of the economy, organizations continue to move beyond national borders. As a consequence, it is also necessary for organizations to view the labor market in an international scope and to select people in an international labor market. If both low-fidelity and high-fidelity simulations are used in international selection practice, a critical research area is the cross-cultural transportability of these simulations. That is, can simulations developed in one culture be transported to and used as a valid predictor in another culture? Although some factors have been identified (see Briscoe, 1997; Lievens, 2006), empirical research is very scarce.

On a general level, it should be noted that simulations are contextualized selection procedures because they are embedded in a particular context or situation. This contextualization makes them particularly prone to cultural differences because the culture in which we live acts like a lens, guiding the interpretation of events and defining appropriate behaviors (Croppanzo, 1998; Lytle, Brett, Barsness, Tinsley, & Janssens, 1995). The stimuli (situations in SJTs, exercises in ACs) presented to the applicants are a first aspect that should be given careful consideration in light of the cross-cultural transportability of simulations. These stimuli are generated from a job analysis (critical incidents). We do not know whether the situations always generalize across cultures. Some situations might simply not be relevant in one culture, whereas they might be very relevant in another culture. Think about the differences in organizing meetings across countries. If we do not take account of these cultural differences, it might well be that applicants are presented with situations (either in AC exercises or SJT item stems) that are simply not relevant in their culture. To our knowledge, no empirical studies have tested whether similar situations are generated across cultures. If simulations want to be truly cross-cultural, it is also important to include cross-cultural situations. This can be easily done in SJTs. In ACs it is also possible. For example, Lievens, Harris, Van Keer, and Bisqueret (2003) developed a cross-cultural AC in which participants of different cultures completed the exercises.

The responses of the candidates and their effectiveness constitute a second facet that might be prone to cultural values and differences. We expect that the range of responses to the situations provided might differ from one culture to another culture. What might be a frequent response (e.g., yelling in a meeting when no one is listening to your opinion) in one culture (e.g., culture low in power distance) might not be endorsed by many applicants in another culture (e.g., culture high in power distance). In addition, we expect that cultural
differences will affect the effectiveness of response options and therefore the evaluation of responses in simulations. For instance, we expect that responses that promote group harmony might be considered more effective in cultures high in collectivism, whereas the reverse might be true in cultures low on individualism.

Therefore, it seems of key importance to tailor the evaluation of responses to the culture of interest. For example, in the cross-cultural AC of Lieveens et al. (2003) people from the host culture (i.e., Japan as selected applicants were required to work in Japan) served as assessors. There is also some evidence of what might happen when the scoring does not match the criterion (culture). For instance, Such and Schmidt (2004) examined the validity of the same SJT in various countries. The SJT was valid in half of the countries, namely the United Kingdom and Australia. Conversely, it was not predictive in Mexico. Thus, the generalizability of simulations such as SJTs and AC exercises to other contexts might be jeopardized if they are scored for a criterion context (e.g., job, organization, culture) different than originally intended. Further research is needed to test the logic of tailoring the scoring key to the host culture as a way of matching predictors and criteria.

A fourth item characteristic that might be prone to cultural differences is the link between responses as indicators for a given construct. Unlike cognitive ability tests, we expect that the item–construct relationship in simulations is more susceptible to deficiency and contamination because of possible cross-cultural differences in the meaning/interpretation of the same situation or same response to the same situation. For example, given the same situation (e.g., a meeting between a supervisor and a group of employees), the same behavior (e.g., clearly and openly defending your views about work standards in front of the supervisor with all employees being present) might be linked to a specific construct (e.g., assertiveness) in one culture (culture low in power distance), whereas it might be an indicator for another construct (e.g., rudeness, impoliteness) in another culture (culture high in power distance).

So far, no studies have explored cultural differences in simulations in terms of the situations, responses, or response–construct linkages. In future research, it might be particularly useful to experiment with a combined eticemic approach (Schmit, Kihm, & Robie, 2000) in developing simulations such as SJTs and AC exercises across cultures. In this approach, items (incidents, responses) are gathered in all cultures in which the simulations will be used. Only those items that are relevant across all cultures are kept. The same strategy of using experts from different cultures might be used for constructing a scoring key. Although such a combined eticemic approach has been successfully applied for constructing a so-called global personality inventory (Schmit et al., 2000), it might be worthwhile to explore its viability for contextualized method-driven selection procedures such as AC exercises and SJT items.

Epilogue

This chapter reviewed high-fidelity and low-fidelity simulations side by side. Our aim was not to demonstrate that one type of simulation is better than the other one. Contrary to such simple examinations of “which simulation is better,” we recommend that researchers focus on the stimulus, response, and scoring features of simulations and examine their effects on selection outcomes such as validity, adverse impact, fakability, and applicant perceptions. When such comparative investigations are informed by a “theory of performance” on simulations, we believe that they show tremendous promise in integrating the research on high-fidelity and low-fidelity simulations instead of treating those two types of simulations as a dichotomy.

References


Fetzer, M. (2011, April). *The latest research on simulations and multimedia SJTs*. Symposium conducted at the 26th Annual Conference of Industrial and Organizational Psychology, Chicago, IL.

Fetzer, M., Tuzinski, K., & Freeman, M. (2010, April). *3D animation, motion capture, and SJTs: I-O is finally catching up with IT*. Paper presented at the 25th Annual Conference of Industrial and Organizational Psychology, Atlanta, GA.


Lievens, F., Buyse, T., & Sackett, P. R. (2005a). The operational validity of a video-based situational judgment test for medical...
college admissions: Illustrating the importance of matching predictor and criterion construct domains. Journal of Applied Psychology, 90, 442–452.


