Journal of Applied Psychology
2011, Vol. 96, No. 5, 927–940

0021-9010/11/$12.00 DOI: 10.1037/a0023496

# The Validity and Incremental Validity of Knowledge Tests, Low-Fidelity Simulations, and High-Fidelity Simulations for Predicting Job Performance in Advanced-Level High-Stakes Selection

Filip Lievens
Ghent University

Fiona Patterson
City University London

In high-stakes selection among candidates with considerable domain-specific knowledge and experience, investigations of whether high-fidelity simulations (assessment centers; ACs) have incremental validity over low-fidelity simulations (situational judgment tests; SJTs) are lacking. Therefore, this article integrates research on the validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations in advanced-level high-stakes settings. A model and hypotheses of how these 3 predictors work in combination to predict job performance were developed. In a sample of 196 applicants, all 3 predictors were significantly related to job performance. Both the SJT and the AC had incremental validity over the knowledge test. Moreover, the AC had incremental validity over the SJT. Model tests showed that the SJT fully mediated the effects of declarative knowledge on job performance, whereas the AC partially mediated the effects of the SJT.

*Keywords:* situational judgment tests, assessment centers, high-stakes selection, medical education

High-stakes testing refers to the use of tests where the test results play a critical role for individuals in getting access to employment, education, or credentialing opportunities (Sackett, Schmitt, Ellingson, & Kabin, 2001). High-stakes testing situations can be further distinguished in entry-level and advanced-level selection situations. Whereas relatively stable abilities and aptitudes are often explicitly assessed in entry-level selection (e.g., admissions testing), a focus on knowledge and skill as products of abilities, traits, and interests is characteristic of most advanced-level selection programs (e.g., selection into advanced training programs, licensure, certification; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999; Raymond, 2001; Rolfhus & Ackerman, 1999; Sackett et al., 2001). Hence, as posited by Raymond, Neustel, and Anderson (2007), advanced-level high-stakes selection instruments typically capture "job-related knowledge while limiting the extent to which general cognitive ability and personality traits influence test scores" (p. 368).

In advanced-level high-stakes testing, various test formats have been used for measuring knowledge and skill. Job knowledge tests represent one common, efficient way of assessing domain-specific knowledge. Simulations constitute a more contextualized albeit more costly approach for evaluating how knowledge and experience acquired have translated into skilled performance. For instance, simulation exercises have been used for certification in legal bar examinations or in teacher assessments (Sackett et al., 2001). In simulations, candidates perform a selected set of tasks that are more or less exact replicas of on-the-job tasks (Roth, Bobko, & McFarland, 2005; Schmitt & Mills, 2001; Thornton & Cleveland, 1990). Hence, simulations have traditionally been categorized as scoring high on fidelity, as they present job-related situations to candidates and require actual behavioral responses from those candidates (Thornton & Rupp, 2006). Examples of high-fidelity simulations are work samples and assessment center (AC) exercises, which are also known in other domains as "performance tests," "performance assessment," or "authentic assessment" (Lane & Stone, 2006; Sackett, 1998).

In recent years, less costly alternatives to the traditional high-fidelity simulations have emerged in the form of situational judgment tests (SJTs). Although not really a new invention (SJTs existed prior to World War II), they were reintroduced by Motowidlo, Dunnette, and Carter (1990), who labeled them "low-fidelity simulations." This term was used because SJTs 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 (McDaniel, Hartman, Whetzel, & Grubb, 2007; Weekley, Ployhart, & Holtz, 2006).

SJTs have gained in popularity in advanced-level high-stakes selection because cost and time constraints often make high-fidelity simulations impractical to develop (e.g., Motowidlo et al., 1990; Patterson, Baron, Carr, Plint, & Lane, 2009). However, the

This article was published Online First April 11, 2011.

Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium; Fiona Patterson, Department of Psychology, City University London, London, England.

We thank Frederik Anseel and George Thornton for their insightful comments on a previous version of this paper. We acknowledge Helen Baron, Victoria Carr, and Bill Irish for their help in preparing materials for this study.

Correspondence concerning this article should be addressed to Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium. E-mail: filip.lievens@ugent.be

lower stimulus and response fidelity of SJTs might also lead to lower criterion-related validity than that of the traditional, more costly high-fidelity simulations. So far, no research has compared the validity of low-fidelity and high-fidelity simulations in advanced-level high-stakes selection within the same setting and sample, as research on these two types of simulations has been conducted independently. From a utility perspective, this means that we do not know whether the less costly low-fidelity simulations (SJTs) also result in lower validity than that of their high-fidelity counterparts (AC exercises). In addition, the incremental validity of low-fidelity and high-fidelity simulations over traditional job knowledge tests and over one another for predicting job performance is not known. From a conceptual perspective, there is also little theory available that might help to explain potential differences in predictive potential of low-fidelity versus high-fidelity simulations in advanced-level high-stakes selection.

Therefore, in this article we use a predictive design to conduct a comparative evaluation of three common predictors (knowledge tests, high-fidelity AC simulations, and low-fidelity SJT simulations) of job performance in an advanced-level high-stakes context. This comparative evaluation is meant to answer the following three questions of central importance in advanced-level high-stakes selection:

1. What is the relative validity of a knowledge test, an SJT (low-fidelity simulation), and an AC (high-fidelity simulation) in predicting job performance?

2. Do an SJT and an AC explain incremental variance in job performance over a knowledge test?

3. Does an AC explain incremental variance in job performance over an SJT?

The next section formulates hypotheses for each of these questions. The hypotheses are then summarized in a theoretical model of how these three common advanced-level selection predictors are related to each other in predicting job performance.

## Study Background

### Validity of Knowledge Tests, Low-Fidelity Simulations, and High-Fidelity Simulations

Knowledge tests, low-fidelity simulations (SJTs), and high-fidelity simulations (ACs) constitute three common procedures for selecting candidates with considerable domain-specific knowledge and experience in high-stakes testing. Knowledge tests are typically used for assessing applicants' declarative knowledge (i.e., knowledge of facts, rules, principles; see Kanfer & Ackerman, 1989; McCloy, Campbell, & Cudeck, 1994). Given that it is uncommon to include a cognitive ability test in advanced-level high-stakes selection (AERA, APA, & NCME, 1999; Raymond et al., 2007), knowledge is often used as a proxy measure for cognitive ability because cognitive ability has been identified as a key determinant of knowledge acquisition and learning (Kanfer & Ackerman, 1989; Ree, Carretta, & Teachout, 1995; Salas & Cannon-Bowers, 2001). Meta-analytic research further documents that knowledge tests are among the best predictors of training and job performance (Schmidt & Hunter, 1998).

Apart from knowledge tests, simulations are routinely used for examining candidates' skills and performance. These simulations traditionally have been high-fidelity simulations in the form of AC exercises or work samples, with meta-analytic research attesting to their good criterion-related validity in employment settings (Arthur, Day, McNelly, & Edens, 2003; Roth et al., 2005). In the last decade, low-fidelity simulations such as SJTs have emerged as alternatives to the traditional high-fidelity simulations. Meta-analyses showed that their reduced fidelity does not seem to jeopardize their criterion-related validity (Christian, Edwards, & Bradley, 2010; McDaniel et al., 2007; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001), although it should be noted that most studies were concurrent in nature and that there were no direct comparisons with high-fidelity simulations.

In sum, meta-analytic research attests to the validity of each of these predictors in employment settings. However, no studies have simultaneously examined their predictive validity relative to each other within the same setting and sample. In addition, joint examinations of these three predictors in an advanced-level high-stakes selection context are lacking. Therefore, this study used a predictive design for investigating the validity of these three predictors in that context. We hypothesized that prior research findings of employment settings would be replicated. Thus,

> *Hypothesis 1a (H1a):* In advanced-level high-stakes selection, knowledge tests will be a significant predictor of job performance.
>
> *Hypothesis 1b (H1b):* In advanced-level high-stakes selection, high-fidelity simulations (ACs) will be a significant predictor of job performance.
>
> *Hypothesis 1c (H1c):* In advanced-level high-stakes selection, low-fidelity simulations (SJTs) will be a significant predictor of job performance.

### Incremental Validity of Low-Fidelity and High-Fidelity Simulations Above Knowledge Tests

It is pivotal to examine the validity of different predictors over and above each other; this process is typically referred to as incremental validity (Schmidt & Hunter, 1998). From a utility standpoint, the use of additional predictors is of value only when they explain variance in the criterion beyond that which is accounted for by other predictors. This aim of broadening "what is being measured" has been one of the main motivations for investing in simulations in high-stakes settings (together with their lower adverse impact; Sackett et al., 2001).

There are various reasons why high-fidelity simulations might offer incremental variance over knowledge tests. One reason is that high-fidelity simulations use generic exercises to gather samples of candidate behavior in job-related situations. Accordingly, they build only to some extent on job-specific knowledge that candidates have already acquired (Sackett et al., 2001; Whetzel & McDaniel, 2009). The rationale is that AC exercises are too expensive to capture only job knowledge. Therefore, the more generic nature of AC exercises should serve as a platform for observing verbal and nonverbal behavior related to knowledge, skills, and abilities (KSAs) outside the cognitive realm. In addition, high-fidelity simulations differ from knowledge tests in terms of their method of measurement: High-fidelity simulations typically have an open-ended response method, whereas knowledge tests use a paper-and-pencil, multiple-choice response method. In light of the above arguments, we hypothesized

> *Hypothesis 2a (H2a):* In advanced-level high-stakes selection, high-fidelity simulations (ACs) will have incremental validity above knowledge tests for predicting job performance.

As noted above, low-fidelity simulations (SJTs) have emerged in recent years as a cost-efficient alternative to high-fidelity simulations. However, due to the lower fidelity of SJTs, questions might be raised as to whether they still produce incremental validity over knowledge tests. In fact, in SJTs and knowledge tests a paper-and-pencil format with multiple-choice questions is typically used. In addition, in order to be seen as realistic and job related, most low-fidelity simulations (SJTs) for selecting candidates with substantial professional experience build more on applicants' existing foundation of knowledge than do high-fidelity simulations (Sackett et al., 2001; Whetzel & McDaniel, 2009). For instance, an SJT item that asks whether a physician will give in to the patient's refusal to take a specific medicine will build on applicants' declarative knowledge (e.g., facts about the disease, medicine).

Despite this potential bigger overlap between low-fidelity simulations (SJTs) and knowledge tests in advanced-level selection, we expected that SJTs would add incremental portions of variance above knowledge tests. This expectation is grounded on the theory of knowledge determinants underlying performance on SJTs developed by Motowidlo and colleagues (Motowidlo & Beier, 2010; Motowidlo, Hooper, & Jackson, 2006, p. 749). In several studies, Motowidlo posited that low-fidelity simulations (SJTs) measure procedural knowledge. That is, apart from assessing job-specific knowledge, SJTs assess the extent of knowledge somebody has acquired about effective and ineffective courses of action in job-related situations such as those described in an SJT. As an SJT entails a variety of situations, SJTs might measure various kinds of procedural knowledge (e.g., how to deal with interpersonal situations, decision-making situations). Consider the example item above about refusal to take medicine. When applicants decide on the basis of declarative knowledge whether they can give in to the patient's request, they are required to use procedural knowledge on how to convey their decision to the patient (see also the example SJT item in Appendix B). So, procedural knowledge about the costs and benefits of engaging in specific trait-relevant behavior is necessary (Motowidlo & Beier, 2010; Motowidlo et al., 2006).

In short, on the basis of these conceptual arguments we expected that in advanced-level high-stakes selection, SJT scores would be related to scores on knowledge tests, as SJTs typically build on candidates' declarative knowledge acquired. As SJTs also aim to capture procedural knowledge about effective behaviors in job-related situations, we further expected that in this context SJTs would explain additional variance over knowledge tests. Thus,

> *Hypothesis 2b (H2b):* In advanced-level high-stakes selection, low-fidelity simulations (SJTs) will have incremental validity above knowledge tests for predicting job performance.

## Incremental Validity of High-Fidelity Simulations Above Low-Fidelity Simulations

The incremental validity issue becomes even more important when one contrasts low-fidelity to high-fidelity simulations. As noted above, high-fidelity simulations are costly to develop, administer, and score, whereas low-fidelity simulations are more easily developed and enable quick administration for screening a large number of applicants. This raises the key question of whether the increased efficiency and lower costs of low-fidelity simulations also come with a "cost" in terms of reduced predictive potential. In other words, do high-fidelity simulations offer incremental validity over their low-fidelity counterparts in predicting job performance in advanced-level high-stakes selection?

This key question has so far remained unanswered because simultaneous examinations of the validity of high-fidelity simulations (AC exercises) and low-fidelity simulations (SJTs) in the same sample are sorely lacking. From a utility perspective, this means that organizations currently are left in the dark as to whether there are any predictive gains of using high-fidelity simulations as supplements to low-fidelity simulations. Alternatively, organizations do not know whether there are any predictive validity losses when relying only on low-fidelity simulations in high-stakes testing.

In hypothesizing about the reasons why high-fidelity simulations might explain incremental variance over low-fidelity simulations for predicting job performance, it is useful to make a distinction between the method and content of measurement (Arthur & Villado, 2008; Hunter & Hunter, 1984). Regarding the method of measurement, AC exercises and SJTs differ in terms of the fidelity with which they present stimuli and capture responses. Fidelity denotes the extent to which the assessment task and context mirror those actually present on the job (Callinan & Robertson, 2000; Goldstein, Zedeck, & Schneider, 1993). In AC exercises, candidates are confronted with actual task stimuli. They might also interact with other candidates and/or role-players. Response fidelity is also enhanced because candidates have to generate solutions for the problems themselves (open-ended behavioral response mode). Conversely, most low-fidelity simulations rely on a paper-and-pencil format wherein candidates are presented with written situations and have to choose what to do or say from a list of predetermined and cued alternatives (close-ended task).

Regarding the content of what is being measured, both high-fidelity and low-fidelity simulations are based on the notion of behavioral consistency: They are based on the assumption that candidates' performance in the simulation will be consistent with candidates' work performance (Motowidlo et al., 1990; Thornton & Cleveland, 1990; Wernimont & Campbell, 1968). However, this behavioral consistency notion is conceptualized differently in high-fidelity and low-fidelity simulations. Low-fidelity simulations (SJTs) assess applicants' procedural knowledge about effective and ineffective courses of behavior in job-related situations, as such knowledge is assumed to be a precursor of effective job behavior (Motowidlo & Beier, 2010; Motowidlo et al., 2006). Conversely, high-fidelity simulations (AC exercises) provide candidates with a platform from which to translate their procedural knowledge and acquired skills into actual behavior, which is then assumed to be consistent with their on-the-job behavior (International Task Force on Assessment Center Guidelines, 2009).

Summarizing this difference between the two types of simulations, Thornton and Rupp (2006) posited that low-fidelity simulations capture procedural knowledge and behavioral intentions, whereas high-fidelity simulations generate behavioral samples (see also Ryan & Greguras, 1998). For example, while SJT items assess whether people know what is the most empathetic, communicative, or resilient option in a given job-related situation, AC exercises examine whether their actual verbal and nonverbal behavioral manifestations are also empathetic, communicative, or resilient. Although candidates might know what the most effective behavioral action is, they might not show this due to their limited behavioral repertoire. Given that procedural knowledge about effective behavior (as measured in SJTs) might exhibit a lower point-to-point correspondence with on-the-job behavior than with actual behavior (as rated in AC exercises), we expected that AC exercises would offer incremental variance over SJTs. Thus, in light of these method and content differences between high-fidelity and low-fidelity simulations, we hypothesized the following:

> *Hypothesis 3 (H3):* In advanced-level high-stakes selection, high-fidelity simulations (ACs) will have incremental validity over and above low-fidelity simulations (SJTs) for predicting job performance.

## Model

Our hypotheses can also be depicted in a model (see Figure 1) that links the three predictors and shows their relation to job performance. In line with prior meta-analyses, this model posits direct effects of each of the three predictors to job performance. Next, consistent with Motowidlo et al.'s (2006) theory of knowledge determinants of SJT, declarative knowledge is a predictor of procedural knowledge (as measured by SJTs). Declarative and especially procedural knowledge are further expected to predict AC performance. The link between procedural knowledge (as measured by SJTs) and AC performance extrapolates on Motowidlo et al.'s argument that SJTs are predictors of job performance because procedural knowledge of effective behavior might be a precursor of showing that behavior on the job. Therefore, the model proposes that procedural knowledge of effective behavior, as measured by SJTs, will also be a precursor of effective behavior in AC exercises, as those exercises aim to simulate job situations.

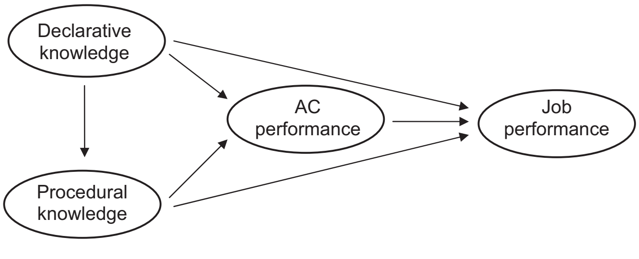

*Figure 1.* Hypothesized model of knowledge determinants of simulations and their link to performance. AC = assessment center.

In sum, the model posits that both declarative knowledge (as measured by knowledge tests) and procedural knowledge (as measured by SJTs) predict performance in high-fidelity simulations (as measured by an AC), which in turn predicts job performance. This model extends Motowidlo et al.'s (2006) theory on knowledge determinants of SJT performance with AC performance and job performance. This integration is conceptually useful because the model shows how these three common, advanced-level selection instruments in combination predict job performance.

# Method

## Sample and Procedure

This study is situated in the context of the selection of general practitioners (i.e., family physicians) in the United Kingdom. This is an advanced-level high-stakes setting because the selection takes place after doctors have undergone 4–6 years of medical education plus at least 2 years of basic medical training. At that point, doctors apply for entry into general practice training via a competitive national selection process (there are typically around three applicants per training place).[1] A similar model of entry into specialty training has been used in Australia and New Zealand (see Jefferis, 2007). Successful doctors must complete up to 3 years of general practice training and then pass certification/licensing examinations to qualify as fully independent (unsupervised) general practitioners. A multistage selection process is used to select which doctors can start training as general practitioners (Patterson et al., 2009).

In 2007, a total of 8,399 candidates went through this multistage selection process at the same time (February–March). This study's sample ($N$ = 500) consisted of general practice trainees who had completed a general practice placement during their first 12 months of training in five randomly chosen UK regions. Trainees who met these inclusion criteria were invited to participate in the study by the UK General Practice National Recruitment Office, resulting in 196 usable responses (response rate of 39.2%). The characteristics of the candidates included in our sample ($N$ = 196) were as follows (the characteristics of the candidate population, $N$ = 8,399, in the 2007 selection are given in parentheses): 40% (48%) men and 60% (52%) women; median age range was 30 years or under (30 or under). According to applicant self-description, 48% (33%) were White, 41% (47%) were Asian, 4% (8%) were Black, and 7% (12%) were from other ethnic groups. Seventy-five percent (51%) of applicants had completed their medical education in the United Kingdom; 79% (71%) had applied for the basic entry level into general practitioner training, and 21% (29%) had applied for higher entry levels. Thus, our sample consisted of slightly more female and White applicants who had completed their education in the United Kingdom. As noted below, our analyses take these small differences into account because we correct for range restriction using data available for the full initial applicant population ($N$ = 8,399) in 2007.

[1] Doctors who are not selected into a general practice training post are still employed as doctors in the UK National Health Service but on service grades rather than in a post that leads to a senior medical appointment.

## Job Analysis

A multimethod job analysis was conducted to define a comprehensive model of performance requirements for the job role of general practitioner (Patterson et al., 2000). In particular, three independent approaches were used: (a) critical incidents focus groups with experienced general practitioners ($N = 35$); (b) behavioral observation and coding of general practitioner–patient consultations ($N = 33$); and (c) critical incidents interviews with patients ($N = 21$). The data triangulated across these three methods led to 11 general practitioner performance dimensions, with associated behavioral indicators. In consultation with an expert panel (comprising three general practitioner training directors), six of the 11 competency domains were judged most critical at the point of selection and were targeted in the selection procedures: empathy, communication, problem solving, professional integrity, coping with pressure, and clinical expertise. Descriptions of each of these six performance dimensions are presented in Appendix A.

## Predictor Measures

**SJT.** The written SJT was developed in accordance with an approach similar to that outlined in other studies (Weekley et al., 2006) and is described in detail in Patterson et al. (2009). Twenty subject matter experts (SMEs; 16 men, 4 women; 1 from a minority ethnic background), who were senior general practitioners with over 10 years of experience of training general practitioners and responsibility for selection within a UK region, attended 2 days of item writing training and worked with three psychologists to generate, review, and refine a bank of items. To this end, the following five performance dimensions were targeted: communication, empathy, professional integrity, coping with pressure, and problem solving (see Appendix A). A knowledge-based response instruction format was chosen, because this format has been found to be less prone to faking (Lievens, Sackett, & Buyse, 2009; McDaniel et al., 2007). Candidates were asked to rank order response options from most to least appropriate or to choose multiple appropriate responses. A total of 186 items was originally written, and these items underwent an extensive cycle of reviews and iterations by SMEs and psychologists. These items were piloted in four test versions in 2006 ($N = 2,552$), and 71% of the items were of sufficient psychometric quality to be included in the final SJT.

A concordance analysis was undertaken to ensure SMEs were in agreement over the scoring key per item. Ten SMEs (senior item writers from the UK Royal College examiner panel for the certification examination), with no previous involvement in the SJT development process, each completed two of four pilot forms to provide a five-person concordance sample. Kendall's *W,* computed for each ranking item, showed 85% concordance over .6 and 71% concordance above .7, indicating adequate interrater agreement. Items with poor concordance were reviewed by the item-writing group and were included in the item bank only if consensus could be reached on the scoring key. Ranking items were scored according to how closely the candidate replicated the scoring key (maximum 20 points per item). For multiple response items, candidates received points for each correct option they chose (maximum 12 points per item). The operational SJT contained 50 items and had a time limit of 90 min. Appendix B presents an example SJT item.

Scores on items that were constructed to capture the same job performance dimensions (on the basis of SME judgments) were averaged to produce dimension scores. However, these SJT dimension scores had low internal consistency reliabilities (range .35–.64). These results are consistent with prior research on SJTs (e.g., Motowidlo, Crook, Kell, & Naemi, 2009; Schmitt & Chan, 2006; Whetzel & McDaniel, 2009) documenting that SJT items are typically construct heterogeneous at the item level. That is, SJT items reflect a variety of performance dimensions. Hence, in our analyses we report at the composite SJT score level. This composite SJT score should be viewed as indicative of a candidate's score on a low-fidelity sample of the job performance domain of interest.

In terms of construct-related validity evidence, prior research showed that this SJT significantly correlated with a biodata inventory measuring the same job performance dimensions ($r = .41$, $p < .001$; Patterson et al., 2009).

**AC.** Each AC lasted 1 day and was typically attended by 48 candidates. Data were collected from approximately 150 ACs over a 6-week period. The AC targeted the same five performance dimensions as the SJT did (see Appendix A). Three work-related simulation exercises (each lasting 20–40 min) were included to cover those dimensions. Exercises and specific exercise content were devised on the basis of job analysis information and input from SMEs. The first exercise was a simulated consultation, in which the candidate took the role of doctor and a medical actor played a patient in a given scenario. The second exercise was a group discussion exercise, in which four candidates were asked to discuss and resolve a work-related issue. The third exercise was a written planning exercise, in which candidates were asked to prioritize a set of impending work-related issues and justify the order chosen. In the AC exercises, candidates were scored on each dimension according to 4-point rating scales (1 = *poor*; 4 = *excellent*) with behavioral anchors. As only one assessor was rating candidate's behavior at any given time, it was not possible to compute interrater reliability. Consistent with current AC practices, every effort was made to have participants rated by different assessors across all exercises. After completion of all exercises, assessors met to discuss their observations and ratings with one another; however, data were integrated with a mechanical procedure allowing for only minor adjustments upon discussion.

AC ratings of the same dimension across exercises were averaged to produce AC dimension ratings. However, these AC dimension scores had low internal consistency reliabilities (e.g., the highest internal consistency was .32). These results are consistent with a large body of research on ACs (for overviews, see Brannick, 2008; Howard, 2008; Lance, 2008). Hence, in our analyses we report at the composite AC score level. This composite AC score should be viewed as indicative of a candidate's score on a high-fidelity sample of the job performance domain of interest.

Experienced general practitioners who supervised general practitioners in training served as assessors. Around eighteen assessors were needed for each AC. All assessors had previously attended a comprehensive training seminar lasting up to one day, led by either a psychologist or a senior general practitioner responsible for selection. Training content was standardized and comprehensive and included the explanation of the specific dimensions and the scenarios in all exercises (information about their role and background information). The training also focused on practice in the process of observing, recording, classifying, rating, integrating, and reporting candidate behavior. Calibration sessions were held, during training and immediately before the AC, to enhance interrater reliability. Assessors also accessed an e-learning resource with 3 hours of instruction containing training materials and practice exercises on observing, recording, classifying, and evaluating candidate behavior using standardized videos with feedback.

**Knowledge test (clinical problem solving).** This test was a machine-graded knowledge test in which candidates applied clinical knowledge to solve a problem reflecting a diagnostic process or to develop a management strategy for a patient (Patterson et al., 2009). This test is primarily a measure of declarative knowledge, covering the range of clinical areas defined by the UK training curriculum. A large item bank was developed by SMEs who were trained in item writing (senior general practitioner trainers with over 10 years' experience of training GPs and responsibility for selection within a UK region) under the supervision of three psychologists. All items were piloted with large samples ($N >$ 400). Candidates scored one point for each correct response identified in multiple best answer items. The final test contained 98 items (with a completion time of 90 min) and had an internal consistency reliability of .88. Appendix C presents an example item of the knowledge test.

## Criterion Measure

Supervisor ratings of trainee job performance on each of the performance dimensions of Appendix A served as criterion measures. Candidates who were successful in the selection process entered a general practitioner training program up to 3 years in duration. During the program, trainees worked under supervision in a number of placements, both in hospitals and in general practice. All the trainees included in the sample had completed a general practice placement during their first 12 months of general practitioner training whereby they were responsible for patients. All trainees began working as general practitioner trainees at the same time and were rated in practice after approximately twelve months in training.

The evaluations were completed by the trainee's general practitioner supervisor, who had directly supervised the trainee in practice and who had met regularly with the trainee to discuss his or her progress. A total of 159 supervisors completed ratings. All supervisors were qualified and experienced general practitioners who had been approved as general practitioner trainers with responsibility for supervising trainees. None of the supervisors had access to the trainees' selection scores when making their assessments. It was made clear that the data gathered would be kept confidential and would be used for research purposes only.

As the above description refers to participants as trainees, a question arises as to whether this evaluation should be viewed as a measure of training performance rather than job performance. We view this as job performance, in that these medical school graduates are engaged in full-time practice of medicine. They are responsible for patients and are working under supervision of a senior general practitioner charged with monitoring and evaluating their work.

The supervisors evaluated performance on each of the job performance dimensions using a 24-item inventory. This inventory consisted of behavioral items associated with each of the job performance dimensions and was developed from an inventory used in a previous validation study (Patterson, Ferguson, Norfolk, & Lane, 2005). A panel of five organizational psychologists selected and agreed on the most representative indicators for each dimension from a bank of possible items. Ratings were made on a 6-point Likert-type scale (1 = *needing significant development* to 6 = *clearly demonstrated*). After the indicators per dimension were averaged, these dimension ratings served as input for a confirmatory factor analysis via EQS (Bentler, 1995). Goodness-of-fit indices showed that a one-factor model produced the best fit to the data, $\chi^2(9) = 64.88$, $p < .00$, Tucker–Lewis index [TLI] = .90, incremental fit index [IFI] = .94, comparative fit index [CFI] = .94, standardized root mean residual [SRMR] = .040.

## Criterion Contamination

In our study, the odds of criterion contamination were slim. Naturally, AC results were used for an initial decision regarding candidates' hiring or rejection. However, these results were not considered subsequently. Moreover, the selection data were not accessible to those who provided the criteria information at the time of assessment. All selection data were confidentially stored in the General Practice National Recruitment Office.

## Range Restriction

For any comparison of selection predictors across different selection stages, careful attention to range restriction is important (Sackett, Borneman, & Connelly, 2008; Sackett & Lievens, 2008; Van Iddekinge & Ployhart, 2008). Along these lines, Sackett et al. (2008) warned,

> Failure to take range restriction into account can dramatically distort research findings. One common scenario where this takes place is in the comparison of the predictive validity of two measures, one of which was used for selection, and hence is range restricted, and the other of which was not. (p. 217)

Indeed, if some predictors are more range restricted than others, the validity of the restricted predictor will be underestimated and the incremental validity of the other predictors will be overestimated.

This admonition also applies to our study because a two-stage selection process was used. First, only candidates who passed the cutoff and top-down selection determined on the basis of a composite of the knowledge test and SJT proceeded to the next stage. A second selection occurred because only candidates who passed the top-down selection determined on the basis of a composite of the various AC dimensions were selected. Given that selection was based on a composite on two occasions, we corrected the correlations for indirect range restriction (Thorndike, 1949, case 3) using the multivariate range restriction formulas of Ree, Carretta, Earles, and Albert (1994). We followed the two-stage approach delineated by Sackett and Yang (2000), began by treating the AC group ($N =$ 196) as the restricted group and the short-listed group ($N = 6{,}542$) as the unrestricted group, and applied the multivariate range restriction formulas to the uncorrected correlations. Next, we treated the short-listed group as the restricted group and the initial applicant pool ($N = 8{,}399$) as the unrestricted group and applied the multivariate range restriction formulas to the theretofore corrected correlations. Statistical significance was determined prior to application of the corrections (Sackett & Yang, 2000).

# Results

## Test of Hypotheses

Table 1 presents the means, standard deviations, and correlations between this study's predictors and overall job performance. The correlation between the SJT and the knowledge test was .50. This was significantly higher than the .30 correlation between the AC and the knowledge test, $t(192) = -2.98$, $p < .01$. This significant difference confirms our expectation that SJTs as measures of procedural knowledge are more related to knowledge tests than is the behavioral format of an AC. The SJT and AC correlated .43. As noted above, this might be due to the fact that SJTs and AC differ not only in terms of method of measurement (paper-and-pencil vs. behavioral response mode) but also in the content of measurement (procedural knowledge about behavioral actions vs. actual behavioral actions).

The first set of hypotheses posited the knowledge test (H1a), the AC high-fidelity simulation (H1b), and the SJT low-fidelity simulation (H1c) to be significant predictors of job performance. The last row in Table 1 presents the correlations between the predictors and the criterion measures. Both uncorrected and range restriction corrected correlations are given. Inspection of the predictor–criterion correlations of Table 1 reveals that all of the three selection procedures emerged as significant predictors of job performance, confirming H1a, H1b, and H1c. None of the corrected predictor–criterion correlations were below .30. Of note, the SJT correlated .37 (corrected $r = .56$) with overall job performance. This was higher than the .30 correlation (corrected $r = .50$) between the AC and overall job performance. Yet, the difference in criterion-related validity between the SJT and AC was not significant, $t(192) = -0.098$, $p = .33$.

The second set of hypotheses dealt with the incremental validity of high-fidelity and low-fidelity simulations over knowledge tests. Therefore, we conducted two incremental validity analyses with the knowledge test entered as the first block. The range-restriction-corrected correlation matrix served as input of those hierarchical regression analyses. Table 2 shows that the AC significantly explained 5.7% of variance above the knowledge test. The other regression analysis revealed that the SJT also significantly explained 5.9% of additional variance above the knowledge test. The additional $R^2$ explained by the SJT and AC respectively over the knowledge test did not differ significantly across the two regressions (Clogg, Petkova, & Haritou, 1995). In sum, these findings lend support to H2a and H2b.

The third hypothesis posited that high-fidelity simulations (ACs) would have incremental validity over and above low-fidelity simulations (SJTs) for predicting job performance. To test this hypothesis, we conducted analyses with and without the knowledge test as control (see Table 2). The AC offered 3.0% of incremental variance over the SJT. In the regression analysis wherein the knowledge test was entered as a control, the incremental variance explained by the AC was 2.1%. The additional $R^2$ explained did not differ across those two regressions. Overall, these results support H3.

## Test of Hypothesized Model

Apart from conducting these incremental validity analyses that are especially relevant from a utility perspective, we sought to examine how these three common predictors work together in combination to predict job performance in advanced-level high-stakes selection. To deepen our understanding of the relationships between these predictors and of the reasons why they predict job performance, we used structural equation modeling (via EQS) to test the model depicted in Figure 1. The range-restricted correlation matrix served as input for these analyses.

Overall goodness-of-fit indices showed that the hypothesized partially mediated model produced a good fit to the data, $\chi^2(99) = 247.59$, $p < .00$, TLI = .94, IFI = .95, CFI = .95, SRMR = .05. Parameter level estimates are presented in Figure 2. Results of the measurement model showed that each of the latent factors was reliably measured, as shown by the high factor loadings associated with the latent procedural knowledge, AC performance, and job performance factors, respectively.[2]

With regard to the structural model, several important findings emerged. First, the path coefficient from declarative knowledge to procedural knowledge was significant, indicating that procedural knowledge as measured by SJTs in this advanced-level context indeed builds on declarative knowledge. Moreover, procedural knowledge seemed to fully mediate the effects of declarative knowledge on job performance, as the direct effect of declarative knowledge on job performance was not significant with the SJT included.

Next, the path coefficient from procedural knowledge to AC performance was significant, whereas the path coefficient from declarative knowledge to AC performance was not. So, in this advanced-level context, AC performance builds less on candidates' extant declarative knowledge base than does the SJT, which was more saturated with this kind of knowledge.

Finally, procedural knowledge as measured with the SJT had not only an indirect effect on job performance through AC performance but also a direct effect on job performance. In addition, even with procedural knowledge included, AC performance continued to have an effect on job performance. So, there was evidence that AC performance partially mediated the effects of procedural knowledge on job performance.

Apart from testing our theoretical model in Figure 1, which is a partially mediated model, we tested two rival models: a direct effects model (only direct effects from each of the three predictors to the criterion) and a fully mediated model (no direct effects from declarative and procedural knowledge to the criterion). The direct effects model produced a bad fit to the data, $\chi^2(100) = 513.86$, $p < .00$, TLI = .82, IFI = .85, CFI = .85, SRMR = .30, whereas the indices of the fully mediated model were indicative of a good fit to the data, $\chi^2(101) = 278.16$, $p < .00$, TLI = .92, IFI = .94, CFI = .94, SRMR = .09. As our theoretical model was nested in this second rival model, it was possible to statistically compare those models. This nested model comparison showed that our hypothesized partially mediated model outperformed the rival fully mediated model, difference in $\chi^2(2) = 30.57$, $p < .01$.

[2] As only a composite score for the knowledge test was available, the factor loading of this test was constrained to 1.00.

Table 1
*Descriptive Statistics and Correlations Between Study Variables (N = 195)*

| Predictor | *M* | *SD* | 1 | 2 | 3 |
|---|---|---|---|---|---|
| 1. Knowledge test | 78.94 | 9.00 | | | |
| 2. Low-fidelity simulation (SJT) | 638.33 | 33.80 | .50 | | |
| 3. High-fidelity simulation (AC) | 3.32 | 0.39 | .30 | .43 | |
| 4. Overall job performance | 4.63 | 0.73 | .36 (.54) | .37 (.56) | .30 (.50) |

*Note*. Correlations in parentheses were corrected for multivariate range restriction. Correlations equal to or above .14 are significant at $p < .05$; correlations equal to or above .19 are significant at $p < .01$. SJT = situational judgment test; AC = assessment center; *SD* = standard deviation.

## Additional Analyses

As noted by an anonymous reviewer, evidence of incremental validity of the high-fidelity approach over the low-fidelity approach is especially important for KSAs that are required at the point of application and are therefore more difficult to train (i.e., noncognitively oriented factors, such as empathy, integrity, coping with pressure, communication) than for more trainable KSAs (i.e., cognitively oriented factors, such as clinical expertise or medical problem solving). Therefore, we also examined whether the incremental validity of ACs over SJTs is different for predicting cognitive versus noncognitive criterion dimensions. To this end, we conducted two separate incremental validity analyses: one for a composite cognitive criterion dimension (a composite of job performance ratings on clinical expertise and problem solving) and one for a composite noncognitive criterion dimension (a composite of job performance ratings on communication, sensitivity/empathy, composite of integrity, and coping with pressure). The knowledge test was always entered as the first block, followed by the SJT as the second block, and the AC as the last block. Results showed that the AC had no incremental validity over the SJT in predicting the cognitive criterion dimension (1.1%, *ns*), whereas it offered incremental validity for the noncognitive criterion dimension (2.5%, $p < .01$).

## Discussion

The use of simulations has a long history in advanced-level high-stakes selection (Lane & Stone, 2006; Sackett et al., 2001; Thornton & Cleveland, 1990). In the past, most emphasis has been put on high-fidelity simulations as supplements for knowledge tests. Work samples and AC exercises (performance assessment) were then deployed for assessing whether candidates could demonstrate that they had acquired the necessary KSAs. Given the surge of interest in less costly low-fidelity simulations, a comparative evaluation of the traditional high-fidelity simulations with their low-fidelity "challengers" is both timely and important. This study addressed this key question using a predictive design in an actual advanced-level high-stakes selection setting with job performance as criterion. Our study provides both practical and theoretical contributions to this domain of selecting candidates with substantial domain-relevant knowledge and experience in high-stakes testing.

### Practical Contributions

Our results revealed that all three predictors (knowledge tests, low-fidelity simulations, and high-fidelity simulations) were valid in this advanced-level selection context. Given the lower costs and

Table 2
*Hierarchical Regressions of Knowledge Test, SJT, and AC for Predicting Overall Job Performance*

| Analysis | Predictor | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|
| H2a | | | | |
| 1. | Knowledge test | .249$^{**}$ | .294 | .294$^{**}$ |
| 2. | High-fidelity simulation (AC) | .197$^{*}$ | .351 | .057$^{**}$ |
| H2b | | | | |
| 1. | Knowledge test | .249$^{**}$ | .294 | .294$^{**}$ |
| 2. | Low-fidelity simulation (SJT) | .250$^{**}$ | .353 | .059$^{**}$ |
| H3 | | | | |
| 1. | Knowledge test | .249$^{**}$ | .294 | .294$^{**}$ |
| 2. | Low-fidelity simulation (SJT) | .250$^{**}$ | .353 | .059$^{**}$ |
| 3. | High-fidelity simulation (AC) | .197$^{*}$ | .374 | .021$^{*}$ |
| H3 | | | | |
| 1. | Low-fidelity simulation (SJT) | .410$^{**}$ | .316 | .316$^{**}$ |
| 2. | High-fidelity simulation (AC) | .231$^{**}$ | .346 | .030$^{**}$ |

*Note*. $N = 195$. Estimates are for last step, not entry. Input correlation matrix was corrected for multivariate range restriction. SJT = situational judgment test; AC = assessment center; H = hypothesis.
$^{*}p < .05$. $^{**}p < .01$.

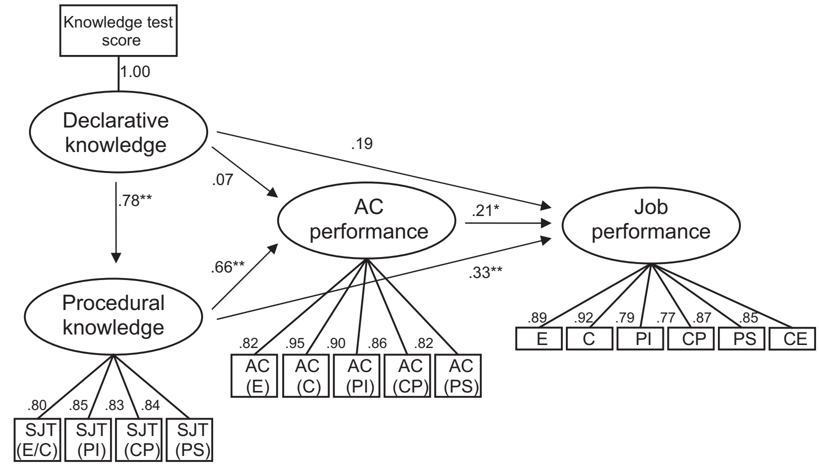

*Figure 2.* Parameter estimates of hypothesized model. AC = assessment center; SJT = situational judgment test; CE = clinical expertise; E = empathy, C = communication; PS = problem solving; PI = professional integrity; CP = coping with pressure.

more efficient administration of low-fidelity simulations, it was of particular interest to examine the validity of low-fidelity versus high-fidelity simulations. In the same advanced-level sample, the validity of the low-fidelity simulations (SJT) was not lower and was even somewhat higher (albeit not significantly) than the validity of high-fidelity simulations (AC). So, in this advanced-level selection context there is evidence that a measure of procedural knowledge produces similar validity coefficients as a measure of actual behavior. These findings are an important contribution to prior research that examined the performance of these two types of simulations separately.

From a utility perspective, the incremental validity analyses further yielded various noteworthy findings. The incremental validity of the SJT over knowledge tests provides a welcome extension over prior findings on SJTs in entry-level high-stakes (admissions) settings that relied on GPA or self-rated/peer-rated criteria (Lievens, Buyse, & Sackett, 2005; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004; Schmitt et al., 2009). Those earlier studies primarily included situations wherein students had little actual experience. So, this study extends the validity of SJTs for predicting supervisory-rated job performance in another high-stakes setting (selection among candidates with considerable domain-specific knowledge and expertise).

The incremental validity analyses also revealed that AC ratings of candidates' behavior in job-related situations had incremental validity for predicting job performance over the SJT as a measure of candidates' procedural knowledge about relevant behavior in written situations. So, organizations can increase the validity of their predictions when adding more expensive high-fidelity simulations to low-fidelity simulations as extra predictive information is garnered when applicants have to translate their procedural knowledge about behavioral actions into actual verbal and nonverbal behavior.

Finally, additional analyses showed that the incremental validity of high-fidelity simulations over low-fidelity simulations for predicting job performance was especially apparent when the job performance criterion was noncognitively oriented instead of cognitively oriented. At a practical level, those incremental validity results are important, as noncognitively oriented dimensions are often more difficult to train than cognitively oriented dimensions (e.g., clinical expertise, medical problem solving). This is especially relevant in the context of selection for high-stakes job roles, such as for physicians.

One explanation for our incremental validity results might be that too few SJT items were considered to tap into these noncognitive skills or that a paper-and-pencil SJT was used. In fact, in the meta-analysis of Christian et al. (2010), the biggest validity difference between paper-and-pencil SJTs (.27) and video-based SJTs (.47) was found for predicting interpersonal skills. So, future research is needed to explore whether incremental validity results similar to those in our study are present for video-based SJTs. As another explanation, the stimulus (i.e., actual interactive social stimuli through role-players, other applicants) and response (i.e., immediate verbal and nonverbal behavioral responses) of high-fidelity simulations might provide those simulations with extra predictive value on candidates' noncognitively oriented performance (Dayan, Kasten, & Fox, 2002; Jansen & Stoop, 2001; Lievens & Sackett, 2006; Olson-Buchanan et al., 1998).

## Theoretical Contributions

Although the incremental validity analyses are especially useful from a practical utility perspective, the test of our model provides interesting theoretical insight and clarification. It also begins to integrate the streams of research on knowledge tests, high-fidelity simulations, and low-fidelity simulations in advanced-level selection by illuminating how each of these predictors works in combination to predict job performance.

Our results revealed that the knowledge determinants of low-fidelity and high-fidelity simulations differed. In this context, high-fidelity simulations (ACs) were not significantly linked to declarative knowledge and therefore seem to represent more generic exercises. Conversely, procedural knowledge as measured by an SJT fully mediated the effects of declarative knowledge on job performance in this advanced-level context. This might be explained by the fact that SJT items in this context build on candidates' declarative knowledge in order to be perceived as realistic. In other words, use of an SJT in an advanced-level context might make the use of a declarative knowledge test redundant. This possibility shows promise for reducing adverse impact against protected groups in selection (Goldstein, Zedeck, & Goldstein, 2002).

Given that the path between declarative knowledge and AC performance was not significant, Figure 2 might be adapted into a cascading model for predicting job performance (i.e., declarative knowledge ⇒ procedural knowledge ⇒ AC performance ⇒ job performance). Note, however, that there is also evidence of a significant direct effect of procedural knowledge (SJT) on job performance. So, AC performance only partially mediated the effects of procedural knowledge as measured by an SJT on job performance, indicating that the inclusion of AC exercises does not make the use of an SJT redundant. This is because in an advanced-level selection context the SJT typically also captures declarative knowledge (contrary to the more generic AC exercises).

On a broader level, results of our model test clarify that one type of simulation is not better than the other one. Contrary to such simple examinations of "which type of simulation is better," our results demonstrate that low-fidelity and high-fidelity simulations supplement each other by capturing different layers of predictive information related to job performance. So, we should not regard low-fidelity and high-fidelity simulations as mutually exclusive.

## Limitations

It is important to note that all conclusions of this study relate to advanced-level high-stakes selection, as this study's context consisted of the selection of general practitioners in the United Kingdom. Although our conclusions might apply to similar high-stakes selection contexts of applicants for professional occupations (e.g., law, business administration, pharmacy) in the profit and public sector, future research should examine the incremental validity of high-fidelity simulations over low-fidelity simulations in entry-level high-stakes selection (admissions settings). Logically, the nomological network will then consist of factors other than declarative (job-specific) knowledge, as ability and personality are likely to be assessed (Raymond et al., 2007).

Other boundary conditions are related to this study's sample. It consisted of experienced applicants with above average cognitive ability who had already passed the hurdles of a competitive educational system. Hence, future research is needed to test our hypotheses in other populations (e.g., customer service occupations).

Finally, our study dealt with an SJT and AC that measured specific performance dimensions. Although the latent factors in our model (declarative knowledge, procedural knowledge, AC performance) are generalizable across similar settings and samples, future research should examine whether the relationships found across these latent factors generalize to SJTs/ACs measuring other dimensions as indicators. This study's SJT had also a knowledge-based response format. Future studies should examine whether our results will translate to SJTs with a behavioral tendency format.

## Directions for Future Research

We envision the following avenues for future research. First, we encourage more integrative work on low-fidelity and high-fidelity simulations. This study took a first step in that direction by testing a model about how simulations are linked to knowledge determinants and performance. Future researchers might extend this model by including cognitive ability and personality determinants (e.g., in the context of entry-level selection).

Second, the use of simulations in high-stakes selection contexts typically has two objectives, namely, broadening the constructs measured and reducing adverse impact. In this study, we focused on the first objective. Future research is needed to ascertain the effects of various low-fidelity versus high-fidelity simulations on adverse impact in high-stakes settings. Such studies might show which specific low-fidelity and high-fidelity simulations provide answers to the validity–diversity trade-off (Ployhart & Holtz, 2008). Along these lines, it might be particularly interesting to investigate hybrids between high-fidelity and low-fidelity simulation, such as SJTs with open-ended response formats or SJTs with behavioral response modes (so-called webcam SJTs).

Third, we need to improve the quality of construct measurement in both low-fidelity (SJT) and high-fidelity (AC) simulations. In the SJT field, recent efforts have been undertaken to develop construct-oriented SJTs (e.g., Bledow & Frese, 2009; Motowidlo et al., 2006). These SJTs are specifically designed to include multiple items for specific constructs. Similarly, it has been suggested in the AC domain that a larger number of shorter AC exercises be used and/or that multiple situational stimuli be deliberately planted within AC exercises (Brannick, 2008; Howard, 2008; Lievens, Tett, & Schleicher, 2009). If such approaches are used, it will be possible to test more fine-grained models in which AC dimension and SJT dimension scores serve as latent factors (instead of as indicator variables).

Future studies should also contrast the coachability of knowledge tests, low-fidelity simulations, and high-fidelity simulations. We know that in the high-stakes-field context of this study, several independent coaching firms helped applicants to be successful in the various tests. So, all predictors of this study were likely to be equally subject to coaching efforts. Laboratory studies (e.g., Cullen, Sackett, & Lievens, 2006), however, might investigate the extent to which knowledge tests, low-fidelity simulations, and high-fidelity simulations are prone to coaching tactics.

In conclusion, this study conducted a comparative evaluation of three common predictors (knowledge tests, high-fidelity simulations, and low-fidelity simulations) in advanced-level high-stakes selection. This study was not about whether one of these predictors outperforms another one. Instead, our results demonstrate that these predictors all build on each other and add different pieces of predictive information for making key selection decisions.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–153. doi:10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93,* 435–442. doi:10.1037/0021-9010.93.2.435

Bentler, P. M. (1995). *EQS: Structural equations program manual.* Encino, CA: Multivariate Software.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62,* 229–258. doi:10.1111/j.1744-6570.2009.01137.x

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 131–133. doi:10.1111/j.1754-9434.2007.00025.x

Callinan, M., & Robertson, I. T. (2000). Work sample testing. *International Journal of Selection and Assessment, 8,* 248–260. doi:10.1111/1468-2389.00154

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63,* 83–117. doi:10.1111/j.1744-6570.2009.01163.x

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology, 100,* 1261–1293. doi:10.1086/230638

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14,* 142–155. doi:10.1111/j.1468-2389.2006.00340.x

Dayan, K., Kasten, R., & Fox, S. (2002). Entry-level police candidate assessment center: An efficient tool or a hammer to kill a fly? *Personnel Psychology, 55,* 827–849. doi:10.1111/j.1744-6570.2002.tb00131.x

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). *g*: Is this your final answer? *Human Performance, 15,* 123–142. doi:10.1207/S15327043HUP1501&02_08

Goldstein, I. L., Zedeck, S., & Schneider, B. (1993). An exploration of the job analysis–content validity process (pp. 2–34). In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 3–34). San Francisco, CA: Jossey-Bass.

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 98–104. doi:10.1111/j.1754-9434.2007.00018.x

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98. doi:10.1037/0033-2909.96.1.72

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17,* 243–253. doi:10.1111/j.1468-2389.2009.00467.x

Jansen, P. G. W., & Stoop, B. A. M. (2001). The dynamics of assessment center validity: Results of a 7-year study. *Journal of Applied Psychology, 86,* 741–753. doi:10.1037/0021-9010.86.4.741

Jefferis, T. (2007). Selection for specialist training: What can we learn from other countries? *BMJ, 334,* 1302–1304. doi:10.1136/bmj.39238.447338.AD

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude–treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74,* 657–690. doi:10.1037/0021-9010.74.4.657

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 84–97. doi:10.1111/j.1754-9434.2007.00017.x

Lane, S., & Stone, C. A. (2006). Performance assessments. In B. Brennan (Ed.), *Educational measurement* (pp. 387–431). Westport, CT: American Council on Education.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). 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. doi:10.1037/0021-9010.90.3.442

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91,* 1181–1188. doi:10.1037/0021-9010.91.5.1181

Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology, 94,* 1095–1101. doi:10.1037/a0014628

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in personnel and human resources management* (pp. 99–152). Bingley, England: JAI Press.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79,* 493–505. doi:10.1037/0021-9010.79.4.493

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91. doi:10.1111/j.1744-6570.2007.00065.x

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740. doi:10.1037/0021-9010.86.4.730

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measured by a situational judgment test. *Journal of Applied Psychology, 95,* 321–333. doi:10.1037/a0017975

Motowidlo, S. J., Crook, A. E., Kell, H. J., & Naemi, B. (2009). Measuring procedural knowledge more simply with a single-response situational judgment test. *Journal of Business and Psychology, 24,* 281–288. doi:10.1007/s10869-009-9106-4

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647. doi:10.1037/0021-9010.75.6.640

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91,* 749–761. doi:10.1037/0021-9010.91.4.749

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51,* 1–24. doi:10.1111/j.1744-6570.1998.tb00714.x

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187–207. doi:10.1037/0021-9010.89.2.187

Patterson, F., Baron, H., Carr, V., Plint, S., & Lane, P. (2009). Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. *Medical Education, 43,* 50–57. doi:10.1111/j.1365-2923.2008.03238.x

Patterson, F., Ferguson, E., Lane, P., Farrell, K., Martlew, J., & Wells, A. (2000). A competency model for general practice: Implications for selection and development. *British Journal of General Practice, 50,* 188–193.

Patterson, F., Ferguson, E., Norfolk, T., & Lane, P. (2005). A new selection system to recruit GP registrars: Preliminary findings from a validation study. *BMJ, 330,* 711–714. doi:10.1136/bmj.330.7493.711

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racio-ethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61,* 153–172. doi:10.1111/j.1744-6570.2008.00109.x

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14,* 369–415. doi:10.1207/S15324818AME1404_4

Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60,* 367–396. doi:10.1111/j.1744-6570.2007.00077.x

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for restriction of range: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79,* 298–301. doi:10.1037/0021-9010.79.2.298

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior knowledge in complex training performance. *Journal of Applied Psychology, 80,* 721–730. doi:10.1037/0021-9010.80.6.721

Rolfhus, E. L., & Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge structures and traits. *Journal of Educational Psychology, 91,* 511–526. doi:10.1037/0022-0663.91.3.511

Roth, P., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58,* 1009–1037. doi:10.1111/j.1744-6570.2005.00714.x

Ryan, A. M., & Greguras, G. J. (1998). Life is not multiple choice: Reactions to the alternatives. In M. Hakel (Ed.), *Beyond multiple-choice: Alternatives to traditional testing* (pp. 183–202). Mahwah, NJ: Erlbaum.

Sackett, P. R. (1998). Performance assessment in education and professional certification: Lessons for personnel selection? In M. D. Hakel (Ed.), *Beyond multiple choice tests: Evaluating alternatives to traditional testing for selection* (pp. 113–129). Mahwah, NJ: Erlbaum.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63,* 215–227. doi:10.1037/0003-066X.63.4.215

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology, 59,* 419–450. doi:10.1146/annurev.psych.59.103006.093716

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56,* 302–318. doi:10.1037/0003-066X.56.4.302

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112–118. doi:10.1037/0021-9010.85.1.112

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52,* 471–499. doi:10.1146/annurev.psych.52.1.471

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274. doi:10.1037/0033-2909.124.2.262

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and practice* (pp. 135–156). Mahwah, NJ: Erlbaum.

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T., Quinn, A., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact of demographic status on admitted students. *Journal of Applied Psychology, 94,* 1479–1497. doi:10.1037/a0016810

Schmitt, N., & Mills, A. E. (2001). Traditional tests and job simulations: Minority and majority performance and test validities. *Journal of Applied Psychology, 86,* 451–458. doi:10.1037/0021-9010.86.3.451

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

Thornton, G. C., III, & Cleveland, J. N. (1990). Developing managerial talent through simulation. *American Psychologist, 45,* 190–199. doi:10.1037/0003-066X.45.2.190

Thornton, G. C., III, & Rupp, D. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development.* Mahwah, NJ: Erlbaum.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61,* 871–925. doi:10.1111/j.1744-6570.2008.00133.x

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests* (pp. 157–182). San Francisco, CA: Jossey-Bass.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372–376. doi:10.1037/h0026244

Whetzel, D. L., & McDaniel, M. A. (2009). Situational judgment tests: An overview of current research. *Human Resource Management Review, 19,* 188–202. doi:10.1016/j.hrmr.2009.03.007

## Appendix A

### Overview of Criterion Dimensions, Predictor Instruments, and Their Linkages

| Criterion dimension | Summary | Example positive indicators | Predictor |
|---|---|---|---|
| Clinical expertise | Capacity to apply sound clinical knowledge and awareness to full investigation of problems. Makes clear, sound, and proactive decisions, reflecting good clinical judgment. | ● Is aware of appropriate clinical options<br>● Shows sound/systematic judgment in making decisions<br>● Is able to anticipate possible issues<br>● Maintains knowledge of current practice | Knowledge test |
| Empathy | Capacity and motivation to take in patient/colleague perspective and sense associated feelings. Generates safe/understanding atmosphere. | ● Responds to patient needs with understanding<br>● Is open, nonjudgmental<br>● Makes efforts to understand patient concerns<br>● Reassures patient<br>● Retains appropriate distance from patient emotions | AC<br>SJT |
| Communication | Capacity to adjust behavior and language (written/spoken) as appropriate to needs of differing situations. Actively and clearly engages patient (and colleagues) in equal/open dialogue. | ● Adjusts response as appropriate<br>● Demonstrates clarity in verbal and written communication<br>● Uses flexible communication style to suit recipients<br>● Establishes equal respect with others | AC<br>SJT |
| Problem solving | Capacity to think/see beyond the obvious, with analytical but flexible mind. Maximizes information and time efficiently and creatively. | ● Thinks "around" issue<br>● Is open to new ideas/possibilities<br>● Generates functional solution<br>● Prioritizes information/time effectively<br>● Is able to identify key points<br>● Sifts peripheral information to detect root cause | AC<br>SJT |
| Professional integrity | Capacity and motivation to take responsibility for own actions (and thus mistakes). Respects/defends contribution and needs of all. | ● Demonstrates respect for patients/colleagues<br>● Is positive when dealing with problems<br>● Is able to admit/learn from mistakes<br>● Is committed to equality of care for all<br>● Puts patient needs before own when appropriate<br>● Backs own judgment appropriately | AC<br>SJT |
| Coping with pressure | Capacity to put difficulties into perspective, retaining control over events. Aware of own strengths/limitations and able to "share the load." | ● Recognizes own limitations<br>● Is able to compromise<br>● Seeks help when necessary<br>● Uses strategies to deal with pressure/stress<br>● Responds quickly and decisively to unexpected circumstances | AC<br>SJT |

*Note.* SJT = situational judgment test; AC = assessment center.

## Appendix B

### Example SJT Item

You are reviewing a routine drug chart for a patient with rheumatoid arthritis during an overnight shift. You notice that your consultant has inappropriately prescribed methotrexate 7.5 mg daily instead of weekly.

Rank in order the following actions in response to this situation (1 = *Most appropriate*; 5 = *Least appropriate*).

a. Ask the nurses if the consultant has made any other drug errors recently.

b. Correct the prescription to 7.5 mg weekly.

c. Leave the prescription unchanged until the consultant ward round the following morning.

d. Phone the consultant at home to ask about changing the prescription.

e. Inform the patient of the error.

(*Appendices continue*)

## Appendix C

### Example of Clinical Problem-Solving Test Items

Reduced Vision

a. Basilar migraine

b. Cerebral tumor

c. Cranial arteritis

d. Macular degeneration

e. Central retinal artery occlusion

f. Central retinal vein occlusion

g. Optic neuritis (demyelinating)

h. Retinal detachment

i. Tobacco optic neuropathy

For each patient below select the SINGLE most likely diagnosis from the list above. Each option may be selected once, more than once, or not at all.

1. A 75-year-old man, who is a heavy smoker, with a blood pressure of 170/105, complains of floaters in the left eye for many months and flashing lights in bright sunlight. He has now noticed a "curtain" across his vision.

2. A 70-year-old woman complains of shadows that sometimes obscure her vision for a few minutes. She has felt unwell recently, with loss of weight and face pain when chewing food.

Received January 22, 2010
Revision received August 31, 2010
Accepted December 10, 2010 ■