Journal of Applied Psychology
2005, Vol. 90, No. 3, 442–452

0021-9010/05/$12.00 DOI: 10.1037/0021-9010.90.3.442

The Operational Validity of a Video-Based Situational Judgment Test for Medical College Admissions: Illustrating the Importance of Matching Predictor and Criterion Construct Domains

Filip Lievens and Tine Buyse
Ghent University

Paul R. Sackett
University of Minnesota

This study is part of a trend of examining noncognitive predictors, for example, a situational judgment test (SJT), as supplements to cognitive predictors for making college admission decisions. The authors examined criterion data over multiple academic years and universities. The criterion domain was broadly conceptualized, including both cognitive and interpersonal domains. The sample consisted of 7,197 candidates of the Medical and Dental Studies Admission Exam in Belgium. Results confirmed the importance of cognitive predictors. A video-based SJT was differentially valid for predicting overall grade point average for different curricula. The SJT showed incremental validity over cognitively oriented measures for curricula that included interpersonal courses, but not for other curricula. The SJT became more valid through the years. This demonstrates the importance of carefully specifying predictor–criterion linkages and of differentiating both predictor and criterion constructs.

The domains of personnel selection and higher educational admissions share a number of parallels (Sackett, Schmitt, Ellingson, & Kabin, 2001). First, in both domains, there is a history of use of cognitively oriented measures of ability and achievement (e.g., cognitive ability and job knowledge tests in employment settings, and measures such as the Scholastic Aptitude Test (SAT), American College Test (ACT), and Graduate Record Examination (GRE) in educational settings).

Second, in both domains, there is substantial current interest in exploring possible supplemental predictors, particularly those outside the cognitive domain. This is motivated by at least two factors. One is the desire to identify a selection system with smaller mean differences by race and, hence, less adverse impact than systems relying heavily or exclusively on cognitive measures. Supplementing cognitive with alternative predictors is seen as a mechanism for accomplishing this. Over the years, a wide variety of alternatives have been proposed, ranging from measuring different constructs (personality or interpersonal skills), using alternative presentation formats, and applying different weighing schemes (cf. Sackett et al., 2001). More unstructured examples of attempting to measure noncognitive factors are letters of recommendation, personal statements, or references. The other motivating factor is a concern about broadening the criterion domain. In employment settings, there has been much recent work on expanded taxonomies of performance, moving beyond task performance to include domains such as citizenship and counterproductive work behavior (Borman & Motowidlo, 1993; Campbell, McCloy, Oppler, & Sager, 1993; Rotundo & Sackett, 2002). In the educational domain, grade point average has been the most common criterion measure, but many colleges and universities also define student success more broadly, including social skill, citizenship, and lifelong learning orientation among many others (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004).

Third, in both domains there has been growing recent interest in one particular type of predictor, namely, the situational judgment test (SJT). In SJTs, applicants are presented with written or video-based depictions of hypothetical scenarios and asked to identify an appropriate response from a list of alternatives (Motowidlo, Dunnette, & Carter, 1990; Weekley & Jones, 1999). In the employment arena, SJTs are becoming increasingly popular for various reasons. First, large-scale studies have shown that SJTs have significant criterion-related validities (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001) and have incremental validity over and above cognitive ability and personality tests (Chan & Schmitt, 2002; Clevenger, Pereira, Wiechmann, Schmitt, & Schmidt-Harvey, 2001). Second, people respond favorably to SJTs because they perceive SJTs to be job-related. Third, SJTs show less adverse impact against minorities than traditional cognitive ability tests (Clevenger et al., 2001). It is important to note, though, that the vast majority of studies are concurrent in design and do not involve the use of SJTs in operational settings (i.e., SJTs were not used for making actual selection decisions, and no predictive validation design was used). In fact, in the meta-analysis of McDaniel et al. (2001), only 6 studies used a predictive validation design, whereas 96 used a concurrent design.

Paralleling these developments is similar work in the educational arena. Initial evidence for the use of SJTs in predicting student performance is encouraging. Hedlund et al. (2001) revealed that an SJT (i.e., tacit knowledge inventory) had incremen-

Filip Lievens and Tine Buyse, Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium; Paul R. Sackett, Department of Psychology, University of Minnesota.

A draft of this article was presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, Illinois, April 2004. We acknowledge Steve Motowidlo for his valuable suggestions on a draft of this article.

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

tal validity over Graduate Management Admission Test scores among Masters of Business Administration students. Likewise, Oswald et al. (2004) constructed an SJT for predicting student performance and examined the validity in terms of predicting both cognitive (grade point average; GPA) and interpersonal domains of student performance in the first year. Their study revealed that the SJT had incremental validity over cognitively oriented measures (ACT or SAT) in predicting a series of performance dimensions related to student college success. Yet, the SJT failed to predict college GPA. Racial subgroup mean differences were much smaller on the SJT than they were on the standardized tests and GPA. It is important to note that these results were obtained in a research setting, which typically lacks the motivational and self-presentational issues inherent in actual high-stakes test programs.

That the SJT of Oswald et al. (2004) predicted performance on dimensions such as leadership and perseverance but did not predict GPA is a useful lead-in to a discussion of an important development in the field, namely, the movement away from general discussions of predictors as valid to consideration of "valid for what?". The taxonomic work on the dimensionality of performance led by Campbell et al. (1993) and illustrated empirically by the U.S. Army's Project A (Campbell et al., 1993) has resulted in more nuanced questions about predictor–criterion relationships. Project A illustrated, for example, that whereas cognitive measures were the most valid predictors of task performance, personality measures were the best predictors of an effort and leadership dimension and a counterproductive behavior dimension (labeled *maintaining personal discipline*; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). Subsequent work has documented the effects of different weightings of criterion dimensions in creating an overall performance measure on the criterion-related validity of various predictors (De Corte, 1999; Hattrup, Rock, & Scalia, 1997; Murphy & Shiarella, 1997). This literature suggests a need to attend to the constructs underlying both predictors and criterion dimensions in developing hypotheses about predictor–criterion relationships. This has direct relevance for SJTs. First, because the SJT approach is a measurement method that can be used to tap various constructs, it is useful to attend to the construct domain underlying a given SJT rather than viewing SJTs as a single entity. Second, expectations for the validity of a given SJT require attention to the criterion in question. An SJT with an interpersonal skills focus would not be theoretically expected to predict a highly cognitively loaded criterion.

Present Study and Hypotheses

Against this backdrop, in this study, we examine the use of traditional cognitive predictors and two alternatives, including an SJT, in the context of admissions for medical and dental studies in Belgium. One difference from admission practices in the United States is that the process in Belgium is centralized and government-run. All students interested in medical and dental studies take an examination battery. Those who pass receive a certificate that permits entry into any of the six medical schools in Belgium. Thus, individual medical schools are not involved in the screening of candidates. This also means that the level of selectivity in Belgium is generally less strict than the level of selectivity in some U.S. medical schools. A second difference is that students enter medical and dental studies at a younger age (e.g., about 19 years of age) rather than on completion of an undergraduate degree, as is more typical in the United States.

A new admissions process was put into place in 1999. It included traditional measures of ability (general cognitive ability) and achievement (science knowledge), as well as two additional predictors: a work sample involving reading a medical article and answering questions about it and a video-based SJT aimed at interpersonal skills and focusing on doctor–patient interactions. We followed four examinee cohorts (1999–2002), monitoring medical school performance through the first 4 years of medical studies.

A crucial feature of the study is that medical schools in Belgium differ in their orientation. These differences stem from strategic choices that the medical schools made to differentiate themselves from one another. One subgroup of medical schools focuses heavily on coursework in the sciences and medical subjects, with little formal attention given to interpersonal aspects of medical practice. Another subgroup also emphasizes sciences and medical subjects, but gives substantial formal attention to interpersonal coursework, starting in the first year and increasing in subsequent years. Thus, whereas both groups of medical schools train students for the same purpose, the formal emphasis placed on interpersonal aspects of medical practice differs across the sets of medical schools. In accordance, grade point average represents a different mix of constructs in these two sets of schools. We posit that GPA is determined primarily by the cognitive predictors of the exam battery in the first context, and we posit that the interpersonally oriented SJT also contributes to the prediction of GPA in the second context.

We proposed the following specific hypotheses. Hypothesis 1 is related to the cognitively oriented predictors. Traditionally, the selection procedure for admission to medical and dental studies has been based on prior academic achievement (e.g., Green, Peters, & Webster, 1993; McManus, 1982; Montague & Odds, 1990), knowledge of science-related subjects (e.g., Montague & Odds, 1990; Tomlinson, Clack, Pettingale, Anderson, & Ryan, 1977), and cognitive abilities (e.g., Roessler, Lester, Butler, Rankin, & Collins, 1978; Vu, Dawson-Saunders, & Barrows, 1987). In general, these cognitively oriented predictors (e.g., Medical College Admission Test) were good predictors of medical students' academic performance (e.g., Green, Peters, & Webster, 1991; Minnaert, 1996; Mitchell, Haynes, & Koenig, 1994; Powis, 1994). On a more general level, a recent meta-analysis of Kuncel, Hezlett, and Ones (2001) showed that a composite of general measures (e.g., GRE Verbal and Numerical) combined with specific GRE subject-matter tests provided the highest validity in predicting academic performance. All of this led to the following hypothesis.

> *Hypothesis 1:* Cognitively oriented predictors (cognitive ability test and science subject tests) will be significantly related to students' GPA.

Three other hypotheses all flow from the discussion above, reflecting the position that an interpersonally oriented SJT should predict performance in a curriculum that emphasizes interpersonal coursework.

Hypothesis 2: If GPA is not only based on medical science courses, but also on courses about interpersonal skills, an SJT measuring interpersonal skills will be significantly related to students' GPA.

Hypothesis 3: If GPA is not only based on medical science courses, but also on courses about interpersonal skills, an SJT measuring interpersonal skills will explain incremental variance over and above more traditional cognitively oriented predictors.

Hypothesis 4: An SJT measuring interpersonal skills will be significantly related to students' scores on courses about interpersonal skills.

This study contributes to the SJT literature in a number of ways. In it, we examined the incremental validity of an SJT over traditional cognitively oriented measures in an operational student admission context. Recall that prior SJT research in admissions settings was done in research rather than in operational settings, and that the vast majority of SJT research in employment settings was done in nonoperational settings. In this study, we examined criterion data over multiple academic years and across multiple universities. The criterion domain is broadly conceptualized, including both cognitive and interpersonal domains, permitting us to test our hypotheses as to whether the validity of the SJT depends on the context in which it is used.

Method

Sample

The total sample consisted of 7,197 candidates (2,606 men and 4,591 women) who attended the Medical and Dental Studies Admission Exam in Belgium between 1999 and 2003. The average age of the candidates was 18 years and 11 months. On average, the passing rate of the admission exam was about 30%. As already noted, candidates who passed the exam received a certificate that warranted entry in any medical university. So, there was no further selection on the part of the universities. However, not all students who passed the exam eventually chose to study medicine.

Only participants who had passed the admission exam, started medical and dental studies in one of the six universities in Belgium, and had continued their studies were included. In total, we were able to obtain the first-year GPAs of 1,768 students, the second-year GPA of 1,087 students, the third-year GPA of 676 students, and the fourth-year GPA of 305 students. Student attrition due to failure (especially in the first academic year) is one reason for the reduction in sample size later in the curriculum. However, the main reason for the sample size reduction across the academic years is the availability of criterion data at the time of this study. In fact, whereas criterion data related to the 1999 exam were available for all 4 academic years, data for 3 academic years were available for the 2000 exam, data for 2 years were available for the 2001 exam, and data for only 1 year were available for the 2002 exam. Note that the restricted samples in later curriculum years did not differ significantly from the total sample in terms of gender or age.

Predictor Measures

We gathered the predictors during the actual admission exam. Each year, this exam lasted for a whole day and was centrally administered in a large hall in Brussels, Belgium. In the morning session, candidates completed the four science tests. In the afternoon, they completed the cognitive ability test, the medical text, and the video-based SJT (physician–patient interaction). The following describes the development and content of each of the predictors used.

Science-related test. Each year, a professor in each respective field developed one of four science-related tests (chemistry, physics, mathematics, and biology). Each science test consisted of 10 questions with four possible answers. Candidate medical students had 180 min to solve these questions. Across the exams included in this study, the average internal consistency coefficient of the science test was .76 (computed across all 40 science questions).

Cognitive ability test. This test consisted of 50 items with five possible answers. Each year, these items were randomly selected from a larger item pool. The items were formulated in either verbal, numeric, or figural terms. Hence, this was a broad cognitive ability test that aimed to measure general mental ability. The time limit was 50 min.

Prior research attested to the good reliability and predictive validity of this test for a medical student population (Minnaert, 1996). In particular, Minnaert (1996) reported an internal consistency of .84 and a validity coefficient of .36 for predicting first-year GPA in medical and dental studies. In this study, the average internal consistency coefficient equaled .71. In light of test security, we cannot mention the source of the cognitive ability test. For the same reason, we cannot present sample items. Interested researchers may contact the authors to obtain more information.

Written medical text. This test was specifically developed for the admission exam. The underlying rationale was to ask candidate medical students to read and understand an article with a medical subject matter. Therefore, this medical text can be considered to be a miniaturized sample of tasks that students will encounter in their medical education. Across the years, examples included a text about diabetes, lower back pain, and so forth. Each text was about 10 pages long and was conceived as a regular scientific article with tables and figures. No statistics were included, and all difficult medical words were explained in an endnote. Students had 50 min in which to read the text and answer the 30 questions. All questions were multiple-choice with four possible answers.

Each year, professors developed the text and the accompanying questions using the same procedures. An existing medical text in a popular medical journal or handbook usually served as starting point. Next, a professor in medicine developed a more elaborate version of the original text. Finally, two professors in medicine assisted us in developing a list of relevant questions and response options. Pilot testing of these questions was not possible because of test security reasons. Likewise, it was forbidden by law to discard specific questions afterward on the basis of received applicant data. Across the exams, the average internal consistency coefficient of this test equaled .74.

Video-based SJT. We developed this test specifically for the admission exam. There is an emerging consensus that SJTs are essentially measurement methods that can be designed to measure a variety of constructs (both cognitive and noncognitive, Chan & Schmitt, 1997, 2002; McDaniel et al., 2001; McDaniel & Nguyen, 2001). Our general aim with the SJT was to measure skills other than cognitive ability (i.e., interpersonal and communication skills). Therefore, the SJT consisted of short videotaped vignettes of key interpersonal situations that physicians are likely to encounter with patients. Each year, we built each SJT around a specific patient and theme (e.g., patient with nausea and chest pain). Although the theme differed across exams, we built the same critical interpersonal incidents (e.g., handling complaints of a patient or conveying bad news) into the interactions. We collected these critical incidents from experienced physicians and professors in general medicine. After each critical incident, the scene froze, and medical student candidates received 25 s to answer the question related to the scene presented. In total, the SJT consisted of 30 multiple-choice questions with four possible answers.

Each year, we followed the same approach (see Motowidlo et al., 1990; Weekley & Jones, 1997) for developing the SJT. We wrote vignettes that nested the critical interpersonal incidents into the general theme. Two

professors teaching physicians' consulting practices always tested the vignettes for realism. Next, we derived questions and response options from the vignettes and critical behaviors. Again, pilot testing and calibration of these questions was not possible because of test security reasons. We asked a panel of experts (experienced physicians and professors in general medicine) to develop a scoring rule. Agreement among the experts was generally satisfactory (Cohen's $\kappa > .70$), and discrepancies were resolved on discussion (e.g., by changing the question or response alternatives), leading to the scoring rule. This scoring rule indicated which response alternative was correct for a given situational item. Endorsement of this response alternative gave the student 1 point. It was forbidden by law to use different scoring rules (e.g., penalizing students for choosing an incorrect alternative by giving them −1 point). Each year, we hired semiprofessional actors and videotaped them delivering the scripted performances in a recording studio. To guarantee realism, an experienced physician attended the set.

Across the exams included in this study, the average internal consistency coefficient for the SJT was .39. SJTs typically demonstrate low internal consistency because the situations and response options presented by SJTs are inherently multidimensional (Chan & Schmitt, 1997, 2002; Clause, Mullins, Nee, Pulakos, & Schmitt, 1998; Motowidlo & Tippins, 1993).

Operational composite. This composite was used to make actual admission decisions; it was a weighted sum of each of the aforementioned predictors. Next, a minimal cutoff was determined on this composite. The weights and cutoff score were determined by law.

Criterion Measures

We retrieved grades for the first 4 years of medical study from archival records of all universities in Belgium. As a first broad criterion, we gathered students' GPAs at the end of each year. In Belgium, GPAs are measured on a scale from 0 to 20, with higher scores indicating better grades. GPAs correlated strongly across years, with correlations among GPAs across years varying between .72 and .78. These values are very similar to the values found in a recent meta-analysis about the temporal stability of GPA (Vey et al., 2003).

As GPA is based on a weighted average of a number of different courses, we also retrieved information about student performance on these courses per year. We did this because GPA is an omnibus measure of academic performance that might not reflect variations in course content across universities, years, and professors. Closer inspection of the curricula of the respective universities showed that there were, indeed, variations across universities in terms of both the courses taught and the weights given to these courses in determining GPA (see the *Analyses Within Curricula* section). Note that student performance on courses is also measured on a scale from 0 to 20, with higher scores indicating better performance.

Results

Preliminary Analyses

As we were to test our hypotheses on data accumulated over 5 years (from 1999 to 2003), we began by examining whether the measurement structure underlying the admission exam was invariant across these 5 years. A model with three factors, namely a cognitively oriented factor (including the cognitive ability test and the four science subject tests, see Kuncel et al., 2001), a factor on which the medical text scores loaded, and a factor related to SJT scores provided a very good fit to the data. In particular, we tested a sequence of increasingly more restrictive tests of measurement invariance. As can be seen in Table 1, there was evidence of full measurement invariance across the five examinations because we found factor form, factor loadings, error variances, and factor variances and covariances to be invariant across the examinations. In addition, the fit of the fully constrained model was still very good, relative noncentrality index = .935, comparative fit index = .951, and root-mean-square error of approximation = .029. Therefore, in the remaining analyses, we report the results for these three factors: cognitively oriented test composite, medical text, and SJT.

Although we found the measurement model to be invariant across years, candidate mean scores per test might still differ across years. This is because the items of the admission exam were not identical across years. As noted above, to preserve the integrity and the security of the original test, we developed alternate forms for each year's test. Hence, mean score changes might occur. Likewise, it can be expected that over the years, candidates obtain higher scores because of the increasing amount of test preparation being provided. Thus, we standardized candidates' test scores within each exam. Given differences across universities, we used a similar approach for the criterion data. In particular, we standardized students' GPA and course scores within university and within academic year.

Overall Analyses

In Table 2, we present the means, standard deviations, and correlations among the predictors. This table is based on all applicants who completed the admission tests between 1999 and 2003. As can be seen, the correlations among the three types of tests were small to moderate. The correlation between the cognitive ability test and the SJT was .19, indicating that the SJT was not heavily cognitively loaded. In their meta-analysis, McDaniel et al. (2001) found a mean correlation between cognitive ability and SJTs of .36 (corrected $r = .46$).

Table 1
Tests of Measurement Invariance for Multigroup Three-Factor Model of Admission Test Scores Across 5 Exam Years (N = 6,005)

Invariance test	χ^2	$\Delta\chi^2$	RNI	CFI	ΔCFI	AGFI	RMSEA	90% CI
Equal number of factors	467.81**		.918	.949		.953	.032	.029–.035
Equal factor loadings	469.40**	1.59	.923	.950	−.001	.955	.031	.028–.034
Equal error variances	471.18**	1.78	.929	.950	.000	.958	.030	.027–.032
Equal factor variances/covariances	471.90**	.72	.935	.951	.000	.961	.029	.026–.031

Note. RNI = relative noncentrality index; CFI = comparative fit index; AGFI = adjusted goodness-of-fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval.
** $p < .01$.

Table 2
Means, Standard Deviations, and Intercorrelations Among Predictors in Applicant Group (N = 7,185)

Predictor	*M*	*SD*	1	2	3
1. Cognitive composite	11.35	2.69	—		
2. Written text	15.81	4.87	.38**	—	
3. SJT	18.73	3.11	.19**	.24**	—
4. Operational composite	20.58	5.17	.92**	.49**	.29**

Note. Although all analyses were conducted on standardized scores, we present the raw scores across exams. The maximum score on each test was 30, with the exception of the operational composite (maximum score = 40). SJT = situational judgment test.
** $p < .01$.

Table 3 displays the means, standard deviations, and correlations of both predictors and criteria. Hence, this table is based only on the subset of applicants who successfully passed the admission exam (i.e., scored higher than the cutoff determined on the operational composite) and subsequently undertook medical studies. A comparison of the descriptive statistics related to the predictors in Tables 2 and 3 reveals the degree of indirect range restriction (Thorndike's Case 3) in each predictor due to the fact that the admission decision was made on the basis of a third variable (the operational composite). As noted above, each predictor is weighted differently in the operational composite, resulting in differing degrees of indirect range restriction. Relative to the applicant pool, those selected scored 1.04 *SD* higher on the cognitively oriented test composite, .23 *SD* higher on the written text, and .16 *SD* higher on the SJT. As indirect range restriction is a special case of multivariate range restriction, we applied the multivariate range restriction formulas of Ree, Carretta, Earles, and Albert (1994) to the uncorrected correlation matrix. After correcting the correlations for range restriction (Stauffer & Mendoza, 2001), we also corrected them for unreliability in the criterion. To this end, we used the mean correlation (.75) between GPA in successive years (see Table 3). We determined statistical significance prior to correcting the correlations (Sackett & Yang, 2000). The values below the diagonal of Table 3 represent the uncorrected correlations between the predictors and academic performance, as measured by GPA in the first, second, third, and fourth year, respectively. The values above the diagonal are the corrected correlations.

Table 3 was the basis for testing Hypothesis 1, namely that cognitively oriented predictors would be related to GPA. Results showed that the composite of the cognitively oriented tests correlated significantly and consistently with students' GPA across the four academic years. This is most evident in the first year. The corrected correlation between this cognitively oriented composite and GPA equaled .52. Conversely, the corrected validities of the other tests were .12 (medical text) and .08 (SJT).[1] These results strongly support Hypothesis 1. As the remaining hypotheses involve differing predictions based on type of curriculum, we now turn to separate analyses within curriculum type.

Analyses Within Curricula

According to our hypotheses, respectable validities for measures that aim to assess a variety of noncognitive interpersonal skills (e.g., the SJT used in this exam) can be expected only if the criterion measures also capture noncognitive dimensions. In particular, Hypothesis 2 stated that if GPA is based not only on medical courses but also on courses about interpersonal skills, an SJT measuring interpersonal skills would be significantly related to GPA.

To test this hypothesis, we inspected the curriculum of the four largest universities in our sample (the remaining two universities were removed because they enrolled only a limited number of students). Across these four universities and across the first four academic years, a total of 105 courses were taught. Next, two of us, Filip Lievens and Tine Buyse, scrutinized the content of these courses and independently rated each course on a five-point scale ranging from 1 (*this is a course with virtually no emphasis on teaching interpersonal/communicative skills related to interactions between physicians and patients*) to 5 (*this is a course with a very strong emphasis on teaching interpersonal/communicative skills related to interactions between physicians and patients*). Interrater agreement among the ratings equaled .92 (intraclass correlation 2.1, Shrout & Fleiss, 1979). Discrepancies between our ratings were easily resolved on discussion.

Courses (e.g., clinical and communicative skills or communication) that obtained a rating of 3 or higher were considered to be courses with an emphasis on teaching interpersonal–communicative skills related to interactions between physicians and patients. Inspection of the distribution of these interpersonal courses across universities revealed that the curricula of universities could be distinguished in terms of their emphasis on interpersonally oriented courses in the first four academic years. Specifically, two groups of curricula emerged. In the curricula of two universities, substantial attention was paid to interpersonal courses. Hence, GPA in these two universities was meaningfully determined by interpersonally oriented courses. In particular, closer inspection of the curriculum showed that the weight of interpersonally oriented courses for determining GPA was .05, .11, .22, and .27 in the first, second, third, and fourth year, respectively. In the curriculum of the two remaining universities, interpersonally oriented courses did not play a meaningful formal role in determining GPA (i.e., the weights of interpersonally oriented courses for determining GPA were .00, .00, .05, and .10 in the first, second, third, and fourth years, respectively).

Table 4 presents the correlations between tests and criteria broken down by type of curriculum. In universities not valuing interpersonally oriented courses in determining GPA, corrected validities for the SJT were low and not significant: .03, .07, .01, and .20 in the first, second, third, and fourth years, respectively. Conversely, the SJT emerged as a significant predictor in curricula

[1] Note also the substantial differences between the corrected and uncorrected correlations. For example, the uncorrected correlation between the cognitive composite and the written text in the first year is .01, whereas the corrected correlation is .17. The key to understanding these differences is found in the specific form of indirect range restriction occurring in this setting. Candidates were selected on an operational composite including all predictors. Selecting on a composite creates a particularly interesting pattern of range restriction for the tests making up the composite: the only way someone with a very low score on one predictor can obtain a high enough composite score to be selected is to have a very high score on another predictor.

Table 3
Means, Standard Deviations, and Correlations Among Predictors and Criteria

Predictor/criterion	*M*	*SD*	1	2	3	4	5
			Year 1 (N = 1,768)				
1. Cognitive composite	14.13	1.66	—	.17	.04	.88	.52
2. Written text	16.94	4.50	.01	—	.14	.31	.12
3. SJT	19.23	2.87	−.05*	.12**	—	.16	.08
4. Operational composite	25.09	3.78	.77**	.21**	.10**	—	.51
5. GPA	13.27	2.53	.33**	.03	.03	.31**	—
			Year 2 (N = 1,087)				
1. Cognitive composite	13.88	1.64	—	.14	.01	.87	.46
2. Written text	15.62	4.23	.03	—	.17	.26	.09
3. SJT	19.60	2.87	−.06*	.15**	—	.14	.09
4. Operational composite	24.84	4.03	.77**	.18**	.09**	—	.45
5. GPA	14.07	1.98	.30**	.03	.05	.28**	—
			Year 3 (N = 676)				
1. Cognitive composite	13.56	1.63	—	.37	.14	.97	.55
2. Written text	16.93	4.01	.10*	—	.26	.48	.29
3. SJT	19.98	2.98	−.03	.20**	—	.24	.20
4. Operational composite	25.99	2.29	.91**	.30**	.14**	—	.55
5. GPA	14.21	1.60	.29**	.11**	.11**	.30**	—
			Year 4 (N = 305)				
1. Cognitive composite	13.51	1.49	—	.52	.45	.96	.37
2. Written text	18.56	4.25	.31**	—	.53	.56	.27
3. SJT	21.35	3.09	.26**	.44**	—	.48	.35
4. Operational composite	26.21	2.27	.91**	.38**	.32**	—	.36
5. GPA	14.45	1.80	.21**	.14*	.24**	.20**	—

Note. Although all analyses are conducted on standardized scores, we present the raw scores. The maximum score on each test was 30, with the exception of the operational composite (maximum score = 40) and grade point average (GPA; maximum score = 20). Uncorrected correlations are below the diagonal; corrected correlations are above the diagonal. Correlations were corrected for multivariate range restriction and criterion unreliability. Statistical significance was determined prior to correcting the correlations. SJT = situational judgment test.
* $p < .05$. ** $p < .01$.

that partially determined GPA on the basis of interpersonally oriented courses in the first years. In particular, corrected validities were .12, .14, .40, and .55 in the first, second, third, and fourth years, respectively. For the last 2 years, the difference in the SJT–criterion correlation coefficients across curriculum type was statistically significant ($p < .001$ in the third year and $p < .05$ in the fourth year). All of this provides support for Hypothesis 2.

In Hypothesis 3, we posited that if GPA is not only based on medical courses but also on courses about interpersonal skills, an SJT measuring interpersonal skills would explain incremental variance over traditional cognitively oriented admission tests. To shed light on this hypothesis, we conducted a hierarchical regression analysis within curriculum type. The matrices corrected for multivariate range restriction and criterion unreliability (see correlations above the diagonals in Table 2) served as input for the hierarchical regression analyses. We determined statistical significance prior to applying the corrections (by conducting hierarchical regressions on the uncorrected matrix of correlations). The cognitive test composite was entered as a first block[2] because these tests have traditionally been used in medical admission exams. Next, we entered the medical text in the regression equation. Finally, we entered the SJT. The results are presented in Table 5. Whereas the SJT never accounted for incremental variance in the first curriculum type, it always explained incremental variance in the second curriculum type. Specifically, in universities that valued courses about interpersonal skills in computing GPA, the SJT accounted for 1% additional variance in the first year, 2% in the second year, 6% in the third year, and 7% in the fourth year. In sum, these results support Hypothesis 3.

As a formal test of whether the beta weights associated with the SJT differed significantly across curriculum type, we also con-

[2] In universities that valued interpersonally oriented courses, it is also interesting to examine whether cognitive predictors predict incremental variance over the SJT. Therefore, we ran a hierarchical regression analysis in which the SJT was entered as the first block. Results showed that the cognitive composite always added incremental variance over and above the SJT (10.2%, 11.6%, 16.1%, and 3.3% for the first, second, third, and fourth year, respectively). This is not surprising, because even in universities that valued interpersonally oriented courses, GPA is still predominantly cognitively loaded (as noted above, the weight given to the interpersonally oriented courses in determining GPA is at most .30).

Table 4
Means, Standard Deviations, and Correlations Among Predictors and Criteria Broken Down by Curriculum Type

	Curriculum with a minimal interpersonal skills component							Curriculum with a substantial interpersonal skills component						
Predictor/criterion	*M*	*SD*	1	2	3	4	5	*M*	*SD*	1	2	3	4	5
	Year 1 (N = 822)							Year 1 (N = 714)						
1. Cognitive composite	14.26	1.67	—	.19	.02	.89	.56	14.10	1.67	—	.15	.04	.86	.49
2. Written text	16.99	4.55	.04	—	.07	.31	.14	16.80	4.47	.00	—	.15	.30	.10
3. SJT	19.18	2.81	−.06	.05	—	.12	.03	19.29	2.84	−.06	.12**	—	.17	.12
4. Operational composite	25.23	3.82	.79**	.21**	.08*	—	.57	24.94	3.86	.75**	.20**	.11**	—	.46
5. GPA	13.08	2.90	.35**	.04	−.01	.35**	—	13.36	2.10	.31**	.03	.07	.27**	—
	Year 2 (N = 484)							Year 2 (N = 433)						
1. Cognitive composite	14.01	1.63	—	.17	−.01	.89	.46	13.88	1.68	—	.12	−.01	.83	.48
2. Written text	15.58	4.24	.06	—	.10	.27	.15	15.37	4.14	.02	—	.18	.24	.05
3. SJT	19.56	2.90	−.08	.09*	—	.11	.07	19.74	2.72	−.08	.16**	—	.14	.14
4. Operational composite	25.08	3.99	.80**	.19**	.07	—	.47	24.61	4.23	.72**	.17**	.10*	—	.44
5. GPA	14.26	2.17	.28**	.07	.04	.29**	—	13.86	1.78	.33**	.00	.10*	.28**	—
	Year 3 (N = 308)							Year 3 (N = 256)						
1. Cognitive composite	13.70	1.59	—	.35	−.05	.97	.46	13.49	1.71	—	.40	.19	.97	.70
2. Written text	16.95	3.98	.09	—	.16	.45	.23	16.74	4.02	.12	—	.27	.52	.40
3. SJT	20.02	2.98	−.14*	.16**	—	.05	.01	20.11	2.84	.00	.20**	—	.29	.40
4. Operational composite	26.18	2.25	.91**	.27**	.03	—	.47	25.91	2.38	.91**	.33**	.18**	—	.72
5. GPA	14.20	1.85	.23**	.08	.00	.24**	—	14.12	1.25	.40**	.17**	.27**	.42**	—
	Year 4 (N = 170)							Year 4 (N = 111)						
1. Cognitive composite	13.59	1.46	—	.46	.35	.96	.28	13.50	1.53	—	.57	.54	.97	.52
2. Written text	18.62	4.25	.27**	—	.45	.51	.16	18.40	4.28	.35**	—	.59	.61	.44
3. SJT	21.59	2.89	.19*	.37**	—	.39	.20	21.08	3.11	.33**	.48**	—	.56	.55
4. Operational composite	26.35	2.28	.91**	.34**	.26**	—	.28	26.14	2.22	.91**	.42**	.38**	—	.52
5. GPA	14.43	2.10	.15*	.07	.12	.15*	—	14.46	1.12	.30**	.25**	.38**	.29**	—

Note. Although all analyses were conducted on standardized scores, we present the raw scores across exams. The maximum score on each test was 30, with the exception of the operational composite (maximum score = 40) and grade point average (GPA; maximum score = 20). Uncorrected correlations are below the diagonal; corrected correlations are above the diagonal. Correlations were corrected for multivariate range restriction and criterion unreliability. Statistical significance was determined prior to correcting the correlations. SJT = situational judgment test.
* $p < .05$. ** $p < .01$.

ducted a regression analysis in which data across the two curriculum types were pooled and the curriculum type and the Curriculum Type × SJT interaction were added as predictors. The SJT × Curriculum interaction effect was not significant and explained only an additional 0.02% and 0.01% of the variance in the first and second year, respectively. Yet in the third and fourth year, the Curriculum Type × SJT interaction explained 1.2% and 1.1% of the variance, respectively. Note that the interaction effect was significant in the third year ($p < .01$) but was not significant in the fourth year ($p = .07$). The fact that the difference in beta weights was not significant in the fourth year is probably due to the moderate power (.68) to detect statistically significant moderators in the smaller samples in the fourth year (Aguinis, Pierce, & Stone-Romero, 1994). This is further supported by the amount of variance explained and the bivariate correlations in Table 4. As noted above, the difference between the SJT–GPA correlations in the fourth year across curriculum type was statistically significant ($p < .05$).

As already mentioned, at the time of this study, criterion data for the third and fourth year were available only for applicants of the 1999 and 2000 exams, whereas criterion data for the first and second year were already available for applicants of the 1999, 2000, 2001, and 2002 exams. This leaves open the possibility that the increasing trend in SJT validity in the third and fourth year might not only result from the type of curriculum but also from the analysis of a different applicant group. Therefore, we also ran all of our analyses for the subset of applicants for whom all criterion data were available at the time of this study (i.e., applicants of the exams of 1999 and 2000). Results showed the same increasing trend in SJT (incremental) validities and the same differential validity pattern for the SJT if our analyses in the first and second year were based solely on applicants of the 1999 and 2000 exams.

Analyses Within Courses

In Hypothesis 4, we posited that an SJT measuring interpersonal skills would be significantly related to courses dealing with interpersonal skills. To test this hypothesis, we computed a validity coefficient between the SJT and students' course grades. As there were 105 courses, this yielded 105 validity coefficients for the

Table 5
Summary of Hierarchical Regression Analyses of Predictors on Grade Point Average (GPA) in First 4 Years Broken Down by Curriculum Type

	Curriculum with a minimal interpersonal skills component					Curriculum with a substantial interpersonal skills component				
Predictor	β	t	p	R^2	ΔR^2	β	t	p	R^2	ΔR^2
GPA Year 1										
Cognitive composite	.56	10.60	.00	.32	.32**	.48	8.98	.00	.24	.24**
Written text	.04	.80	.42	.32	.00	.02	.38	.70	.24	.00
SJT	.01	.25	.80	.32	.00	.10	2.47	.01	.25	.01
GPA Year 2										
Cognitive composite	.45	6.42	.00	.21	.21**	.48	7.57	.00	.23	.23**
Written text	.06	1.19	.24	.22	.00	−.03	−.59	.56	.23	.00
SJT	.06	1.22	.22	.22	.00	.15	2.88	.00	.25	.02**
GPA Year 3										
Cognitive composite	.44	4.02	.00	.21	.21**	.62	7.10	.00	.50	.50**
Written text	.07	1.03	.30	.22	.01	.08	1.31	.19	.51	.02*
SJT	.02	.37	.71	.22	.00	.26	4.53	.00	.58	.06**
GPA Year 4										
Cognitive composite	.24	1.73	.08	.08	.08*	.30	1.91	.06	.27	.27**
Written text	−.01	−.10	.92	.08	.00	.06	.48	.64	.30	.03
SJT	.12	1.21	.23	.09	.01	.35	2.89	.00	.38	.07**

Note. The corrected matrices served as input for the regression analysis. We determined statistical significance prior to correcting the correlations (by conducting the same regression analyses on the uncorrected matrices). Parameter estimates are for final step, not entry. Due to rounding, ΔR^2 differs by .01 from the cumulative R^2. SJT = situational judgment test.
* $p < .05$. ** $p < .01$.

SJT. Next, we paired these validity coefficients with the ratings of the interpersonal orientation of each course (see the *Analyses Within Curricula* section) and correlated these validity coefficients with these ratings. In line with Hypothesis 4, the correlation between the SJT validity coefficients and the interpersonal course ratings was positive ($r = .21$, $p < .05$), supporting the idea that SJTs have higher validity in more interpersonally oriented courses.

Discussion

This study is part of a growing trend of examinations of noncognitive predictors as supplements to well-established cognitively oriented predictors. We offered several hypotheses and found support for each of them. First, this study confirmed the importance of cognitively oriented predictors. This demonstrates that alternative measures are not designed to replace the traditional cognitively oriented predictors. Instead, they are meant to increase the coverage of skills not measured by traditional predictors.

Second, a video-administered SJT was found to be differentially valid for predicting overall GPA for different curricula. The SJT exhibited incremental validity over the cognitively oriented predictors for curricula with a substantial interpersonal skills component but not for curricula with a minimal interpersonal skills component. Within curricula, we further found that the SJT was predictive for interpersonal domains. All of this demonstrates the importance of carefully specifying predictor–criterion linkages and of differentiating both predictor and criterion constructs (Murphy & Shiarella, 1997; Reeve & Hakel, 2002; Rothstein, Paunonen, Rush, & King, 1994; Schmitt & Chan, 1998). Thus, as conceptualizations of job performance broaden beyond task performance to include the citizenship and counter productivity domains, it is important for organizations to carefully identify the criterion constructs of interest and to choose potential supplemental predictors on the basis of hypothesized links to these criterion constructs. It is also important to keep in mind the primacy of the criterion. On finding that the SJT does not show incremental validity for curricula with a minimal interpersonal skills component, one reaction might be to call for a broadening of the curriculum to include interpersonal courses, thus making it likely that the SJT would show incremental validity. We argue against this, as it reflects letting an interest in a predictor drive the choice of the criterion dimensions. Assuming a clear decision on the part of these universities to make a strategic choice to differentiate themselves from one another by choosing to either emphasize or deemphasize an interpersonal skills orientation, predictor choice should follow from this.

This study is also one of the first to examine the effectiveness of SJTs in an actual admission context (see also Lievens & Coetsier, 2002). It extends the results of Oswald et al. (2004) that were obtained in a research setting to actual admission decisions and across a longer time period. The positive news is that this study demonstrates that SJTs can be a useful and valid complement to traditional student admission tests even in an operational high-stakes context. This is important news, because experimental research has shown that SJTs might be prone to faking, which is expected to decrease their validity (Haas & McDaniel, 1999; Nguyen, McDaniel, & Biderman, 2002; Peeters & Lievens, 2005). If students attempted to respond in a socially desirable manner on the SJT, as they may have been motivated to do in a setting such as an admission exam, then such attempts do not seem to invalidate the SJT. Research in other domains has found similar divergent results between faking in controlled lab settings and motivated field settings. In fact, a meta-analysis of Edens and Arthur (2000)

confirmed that real-life motivational distortion results in smaller effect sizes than does instructionally induced faking in laboratory studies, suggesting that laboratory findings may be a worst-case scenario in comparison to faking in actual selection situations (e.g., Rosse, Stechner, Levin, & Miller, 1998). As we were not able to distinguish fakers and nonfakers in this field setting, more research is needed in this area.

With this study, we also contribute to the SJT literature by using a predictive validation design and investigating the validity of SJTs in the long run (4 academic years). The meta-analysis of McDaniel et al. (2001) demonstrates that almost all SJTs have been validated using a concurrent design and in the short run. Along these lines, an interesting result was that the SJT validity increased through the academic years. This is consistent with research showing that noncognitive predictors become more important when the criterion data are gathered later on (Goldstein, Zedeck, & Goldstein, 2002; Jansen & Stoop, 2001).

A final interesting finding is that an SJT measuring interpersonal skills in a physician–patient interaction was predictive even though it was administered to students who had never conducted an interview with a patient. This is relevant in the context of the common assumption that SJTs are primarily measuring job knowledge or experience. Clearly, what is involved here is not job-specific knowledge but, rather, more general knowledge of effective behavior in interpersonal settings.

Future Research Directions

Although this study provides encouraging news for the use of alternative predictors such as SJTs in student admissions, future research is needed to examine other potential advantages and disadvantages of these alternative measures. First, research is needed to investigate whether SJTs lead to lower adverse impact. Initial evidence obtained by Oswald et al. (2004) seems to provide evidence that this is the case. In addition, SJT research in personnel selection has generally found lower subgroup differences for SJTs than for cognitive ability tests (Clevenger et al., 2001; Motowidlo et al., 1990; Weekley & Jones, 1997, 1999), even though the occurrence of adverse impact seems to be moderated by the presentation method (Chan & Schmitt, 1997) and the constructs measured by the SJTs (Schmitt, Clause, & Pulakos, 1996). In particular, SJTs with a lower cognitive loading seem to have less adverse impact than SJTs that are more *g*-loaded.

As a second avenue for future research, we need to examine whether SJTs are prone to practice effects. As argued by Sackett (2005), for any new measure to be a useful part of a large-scale testing program, knowledge of the items on an initial form must not materially affect performance on subsequent alternate forms. Equally important, when SJTs become popular in a student admission context, test preparation firms attempt to teach people how to respond to them most effectively. We still do not know whether coaching is a possible threat to the use of SJTs in a student admission context. Thus, future research should examine the effects of practice and coaching on mean SJT scores and validity. We note that in the current context, there was no evidence of systematic efforts to coach candidates on the SJT. The science tests are weighted most heavily in the operational decision process, and test preparation efforts appeared to be focused on this area.

Third, future studies could determine how different modalities of SJTs impact their SJT performance and validities. For example, in this study, we developed a video-based SJT. It remains unknown whether the predictive validity of an expensive video-based SJT is higher than the predictive validity of a less expensive written SJT, holding the content and items constant. Likewise, it would be interesting to manipulate different response scoring instructions. Prior research has shown that knowledge-related instructions (*what is the most effective answer?*) were more faking-resistant, more related to cognitive ability, and more valid than behavioral tendency instructions (*what would you do?*, Nguyen et al., 2002). In the SJT of this study, we used a mixture of knowledge-related and behavioral tendency instructions.

Limitations

This study has several limitations. A first possible limitation is that this study was conducted in Belgium. As mentioned above, there are some differences between admission practices in Belgium versus those in the United States (e.g., level of centralization of admission process, level of selectivity of entry and age of students). Despite these differences, it should also be noted that there are many similarities between medical education in Belgium and medical education in the United States. Most important, in the United States, there is also a trend in medical schools to broaden their curricula with a focus on communication skills in the first year or, at least, in later years of the curriculum (e.g., Blumberg, 2003; Teutsch, 2003).

A second limitation is that we gathered criterion data for only the first 4 years of medical study. For any predictor, it is critical to examine how it correlates with actual job performance. Therefore, it is necessary to collect criterion data for the whole curriculum and eventually for actual physician performance.

Third, some might argue that our treatment of academic school performance (GPA) as a weighted combination of cognitive and interpersonal characteristics has little value. We do not believe that this is the case. In fact, our validity analyses with school performance as a weighted sum of different course grades can be compared with validity analyses with job performance as a weighted combination of task and contextual performance dimensions. Accordingly, our criterion acknowledges the multidimensionality of the criterion domain and the fact that criterion dimensions might be differentially weighted (Murphy & Shiarella, 1997). On a practical level, organizations (but also departments and even supervisors) typically differ in terms of the emphasis they place on task versus contextual performance in their overall job performance ratings (Johnson, 2001; Rotundo & Sackett, 2002). Thus, an omnibus measure such as job performance might have a different meaning from organization (department) to organization (department). Compare, for example, traditional hierarchic organizations (departments) with democratic team-based organizations (departments). In a similar vein, in our study, an omnibus measure such as GPA has a different meaning depending on the group of medical schools under investigation.

Finally, the SJT in this study had a low internal consistency. This might be because the SJT items were scored with either 0 or 1. In general, relatively low internal consistencies seem to be a common finding of most SJTs given their multidimensional nature (McDaniel et al., 2001). Despite this low internal consistency, the

SJT still had substantial validity. Along these lines, one might wonder whether attempts to make SJTs less multidimensional and therefore more internally consistent might lead to decreases in predictive validity. Similar arguments have been made for biodata inventories (Reiter-Palmon & Connelly, 2000). It is clear that future studies should examine how reliability and validity are related in the context of multidimensional measures such as SJTs.

Conclusions

In this study, we aimed to expand the predictor and criterion domain in student admissions. In terms of the predictor domain, we examined the predictive validity of traditional cognitively oriented predictors and alternative predictors (e.g., SJTs) in an actual student admission context. In addition, we used a broad conceptualization of the criterion domain as we focused on both cognitive and interpersonal domains over a 4-year period. Results confirmed the importance of cognitively oriented predictors. Furthermore, the SJT emerged as a valid predictor in curricula that valued interpersonal skills and in interpersonal courses. In addition, the SJT became more valid through the years and provided incremental variance over and above cognitively oriented predictors, indicating that SJTs enable the broadening of the range of skills measured. All of this suggests that SJTs might be a useful and welcome complement to traditional admission tests, though practice and coaching effects remain a key unexplored issue.

References

Aguinis, H., Pierce, C. A., & Stone-Romero, E. F. (1994). Estimating the power to detect dichotomous moderators with moderated multiple regression. *Educational and Psychological Measurement, 54,* 690–692.

Blumberg, P. (2003). Multidimensional outcome considerations in assessing the efficacy of medical educational programs. *Teaching and Learning in Medicine, 15,* 210–214.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco: Jossey-Bass.

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

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233–254.

Clause, C. C., Mullins, M. E., Nee, M. T., Pulakos, E. D., & Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example. *Personnel Psychology, 51,* 193–208.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt-Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86,* 410–417.

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology, 84,* 695–702.

Edens, P. S., & Arthur, W. (2000, April). *A meta-analysis investigating the susceptibility of self-report inventories to distortion.* Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Goldstein, H. W., Zedeck, S., & Goldstein, I. L. (2002). *g*: Is this your final answer. *Human Performance, 15,* 123–142.

Green, A., Peters, T. J., & Webster, D. J. (1991). An assessment of academic performance and personality. *Medical Education, 25,* 343–348.

Green, A., Peters, T. J., & Webster, D. J. (1993). Preclinical progress in relation to personality and academic profiles. *Medical Education, 27,* 137–142.

Haas, A., & McDaniel, M. A. (1999, April). *Faking strategies: Effects on a situational judgment test.* Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Hattrup, K., Rock, J., & Scalia, C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predictor performance. *Journal of Applied Psychology, 82,* 656–664.

Hedlund, J., Plamondon, K., Wilt, J., Nebel, K., Ashford, S., & Sternberg, R. J. (2001, April). Practical intelligence for business: Going beyond the GMAT. In J. Cortina (Chair), *Out with the old, in with the new: Looking above and beyond what we know about cognitive predictors.* Symposium conducted at the 16th Annual Convention of the Society for Industrial and Organizational Psychology, San Diego, CA.

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.

Johnson, J. W. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology, 86,* 984–996.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment, 10,* 245–257.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103–113.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335–354.

McManus, I. C. (1982). A-level grades and medical school admission. *British Medical Journal, 284,* 1654.

Minnaert, A. (1996). *Academic performance, cognition, metacognition and motivation. Assessing freshmen characteristics on task: A validation and replication study in higher education.* Unpublished doctoral dissertation, University of Louvain, Belgium.

Mitchell, K., Haynes, R., & Koenig, J. (1994). Assessing the validity of the updated medical college admission test. *Academic Medicine, 69,* 394–401.

Montague, W., & Odds, F. C. (1990). Academic selection criteria and subsequent performance. *Medical Education, 24,* 44–47.

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.

Motowidlo, S. J., & Tippins, N. (1993). Further studies of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66,* 337–344.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests:

Multivariate framework for studying test validity. *Personnel Psychology, 50,* 823–854.

Nguyen, N. T., McDaniel, M. A., & Biderman, M. D. (2002, April). *Response instructions in situational judgment tests: Effects on faking and construct validity.* Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario.

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.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65,* 70–89.

Powis, D. A. (1994). Selecting medical students. *Medical Education, 28,* 443–469.

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.

Reeve, C. L., & Hakel, M. D. (2002). Asking the right questions about *g*. *Human Performance, 15,* 47–74.

Reiter-Palmon, R., & Connelly, M. S. (2000). Item selection counts: A comparison of empirical key and rational scale validities in theory-based and non-theory-based item pools. *Journal of Applied Psychology, 85,* 143–151.

Roessler, R., Lester, J. W., Butler, W. T., Rankin, B., & Collins, F. (1978). Cognitive and non-cognitive variables in the prediction of preclinical performance. *Journal of Medical Education, 53,* 678–681.

Rosse, J. G., Stechner, M. D., Levin, R. A., & Miller, J. L. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83,* 634–644.

Rothstein, M. G., Paunonen, S. V., Rush, J. C., & King, G. A. (1994). Personality and cognitive ability predictors of performance in graduate business school. *Journal of Educational Psychology, 86,* 516–530.

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy capturing approach. *Journal of Applied Psychology, 87,* 66–80.

Sackett, P. R. (2005). The performance-diversity tradeoff in admissions testing. In W. Camara & E. Kimmel (Eds.), *Choosing students: Higher education admission tools for the 21st century* (pp. 109–125). Mahwah, NJ: Erlbaum.

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.

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

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach.* Thousands Oaks, CA: Sage.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job relevant constructs. *International Review of Industrial and Organizational Psychology, 11,* 115–139.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Stauffer, J. M., & Mendoza, J. L. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika, 66,* 1–6.

Teutsch, C. (2003). Patient–doctor communication. *Medical Clinics of North America, 87,* 1115–1145.

Tomlinson, R. W. S., Clack, G. B., Pettingale, K. W., Anderson, J., & Ryan, K. C. (1977). The relative role of "A" level chemistry, physics and biology in the medical course. *Medical Education, 11,* 103–108.

Vey, M. A., Ones, D. S., Hezlett, S. A., Kuncel, N. R., Vannelli, J. R., Briggs, K. H., & Campbell, J. P. (2003, April). *Relationships among college grade indices: A meta-analysis examining temporal influences.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Vu, N. V., Dawson-Saunders, B., & Barrows, H. S. (1987). Use of medical reasoning aptitude test to help predict performance in medical school. *Journal of Medical Education, 62,* 325–335.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50,* 25–49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679–699.

Received December 19, 2003
Revision received April 29, 2004
Accepted May 4, 2004 ■