Journal of Applied Psychology
2004, Vol. 89, No. 2, 187–207

Copyright 2004 by the American Psychological Association
0021-9010/04/$12.00 DOI: 10.1037/0021-9010.89.2.187

Developing a Biodata Measure and Situational Judgment Inventory as Predictors of College Student Performance

Frederick L. Oswald, Neal Schmitt, Brian H. Kim, Lauren J. Ramsay, and Michael A. Gillespie
Michigan State University

This article describes the development and validation of a biographical data (biodata) measure and situational judgment inventory (SJI) as useful predictors of broadly defined college student performance outcomes. These measures provided incremental validity when considered in combination with standardized college-entrance tests (i.e., SAT/ACT) and a measure of Big Five personality constructs. Racial subgroup mean differences were much smaller on the biodata and SJI measures than on the standardized tests and college grade point average. Female students tended to outperform male students on most predictors and outcomes with the exception of the SAT/ACT. The biodata and SJI measures show promise for student development contexts and for selecting students on a wide range of outcomes with reduced adverse impact.

Within the complex and competitive admissions process, colleges and universities seek out the best students possible for their institutions, in which "best" can be defined in many ways. Traditionally, selection systems for college admissions typically use standardized tests of verbal and mathematical skills, and possibly records of achievement in specific subject matter areas. Such systems have worked well for decades, especially in comparison with alternatives that have been used or considered. Generally speaking, standardized cognitive ability tests are efficient for mass distribution and application, provide a standard of comparison across differing educational backgrounds, and demonstrate largely unparalleled criterion-related validities of approximately $r = .45$ with cumulative college grade point average (GPA), in addition to smaller but practically significant relationships with study habits, persistence, and degree attainment (Hezlett et al., 2001). However, critics argue there is substantial room for improvement with respect to the validity and practical utility of current selection tools (Breland, 1998; Payne, Rapley, & Wells, 1973).

In fact, some individuals such as Atkinson (2001), head of the University of California system, have called for abandoning the SAT-I (test of general verbal reasoning, reading, and math problem-solving skills) and replacing it with a test or tests more closely related to school curricula, like the SAT-II (measures of English, history, science, social studies, math, and languages). Various stakeholders in admissions testing have been more strident in demanding new selection tools with adequate criterion-related validity, less adverse impact, and greater relevance to a broader conceptualization of performance in college. If new selection tools are to improve on the demonstrated criterion-related validity of the current knowledge and cognitively based predictors, it is likely that these tools will need to be based on a clear identification, definition, and measurement of a broader set of performance outcomes (Sackett, Schmitt, Ellingson, & Kabin, 2001) than the GPA and graduation outcomes usually used to evaluate instruments such as the SAT and the American College Testing (ACT) battery and similar instruments. As evidenced in mission statements and other promotional materials in print and on the Internet, colleges clearly want students who will succeed in the college environment, whether that means succeeding academically, interpersonally, psychologically, or otherwise. If one takes seriously what colleges claim to want in their students, then we argue that it is appropriate to reconsider traditional GPA and graduation criteria on several accounts. First, traditional academic outcomes are useful for what they are intended to measure but insufficient when one considers the entire experience contributing to students' performance and success in college. Furthermore, GPA as a composite measure is not standardized and may represent the outcome of some very different student behaviors, as reflected in different types of courses taught by different instructors. For instance, it is likely that students self-select into classes of different difficulty level or content domains on the basis of their ability and interest (Goldman & Slaughter, 1976), and thus student GPAs are not directly comparable.

Borrowing from current theories of job performance (Campbell, Gasser, & Oswald, 1996; Campbell, McCloy, Oppler, & Sager, 1993; Rotundo & Sackett, 2002), we reexamined the domain of college performance. Although traditional criteria for college student performance have tended to fall under the broad category of task performance (e.g., grades on specific assignments, measures of technical knowledge, GPA), the converging themes in college mission statements and other information about higher education encouraged us to expand into a criterion space that captures alternative dimensions such as social responsibility (Borman & Motowidlo, 1993) and adaptability and life skills (Pulakos, Arad, Donovan, & Plamondon, 2000). If the criterion domain of college

Frederick L. Oswald, Neal Schmitt, Brian H. Kim, Lauren J. Ramsay, and Michael A. Gillespie, Department of Psychology, Michigan State University.

We thank Michele Boogaart, Michael Jenneman, Elizabeth Livorine, and Justin Walker for their data collection efforts, as well as the College Board for their support of this project.

Correspondence concerning this article should be addressed to Frederick L. Oswald, Michigan State University, Department of Psychology, 129 Psychology Research Building, East Lansing, MI 48824-1117. E-mail: foswald@msu.edu

performance is in fact broader and more complex than traditionally conceived and measured, then this in turn implicates broader and more complex combinations of individual abilities and characteristics that predict performance. Specifically, with a broader college performance domain against which admissions decisions are validated, we should find that measures of noncognitive constructs, such as social skills, interests, and personality, are also valid predictors of performance in college. Using a broad set of predictors that capture noncognitive as well as cognitive individual characteristics may reduce the level of adverse impact that typically results from the large subgroup mean differences on cognitively based tests (Hough, Oswald, & Ployhart, 2001; Sackett et al., 2001). Combining cognitive predictors with less cognitive alternative predictors in a compensatory model (such as a linear regression model) should then allow for selecting individuals with somewhat lower levels of some cognitive abilities yet overall are still desirable college applicants.

Most of today's colleges typically base applicant selection decisions on some combination of academic records, high school GPA, class rank, and SAT or ACT score (Breland, 1998; McGinty, 1997). Like standardized tests, high school GPA and rank appear to have relatively high criterion-related validities with college GPA, with correlations between .44 and .62 once corrections for measurement unreliability and range restriction are made (Hezlett et al., 2001). The use of other selection tools varies greatly because colleges can choose the "educational criteria, including racial diversity, that they wish to consider in admissions, so long as they do not apply different standards to different groups" (*Regents of the University of California v. Bakke*, 1978). Depending on their selectivity and demographic characteristics (e.g., public or private), colleges often request additional information about an applicants' prior achievements, background experiences, nonacademic talents, and interpersonal skills, all of which are intended to provide a holistic view of applicants and indicate the likelihood of their success in or contribution to a college. Popular methods of obtaining such information include achievement test scores, letters of recommendation, personal statements, lists of extracurricular activities, interviews, and peer references. There exists some support for the incremental validity and practical usefulness of such measures over the more common predictors mentioned above (Cress, Astin, Zimmer-Oster, & Burkhardt, 2001; Ra, 1989; Willingham, 1985). However, these supplementary measures are problematic to the extent that (a) admissions personnel pay attention to, interpret, and weight this information in different ways; (b) admissions personnel rely on information about students' past experiences that is to some extent idiosyncratic and not in a standardized format; (c) collecting and evaluating this information requires extra cost in time and resources; and (d) information is self-reported and may be difficult to verify (Willingham, 1998). Not implementing, scoring, or weighting such measures in a systematic manner across colleges, and not tying these measures to a relatively broad domain of college performance where supplementary measures may be more useful, preclude a solid conceptual understanding and a consistent and practical level of incremental validity above standardized test scores and high school GPA.

This article describes the development and validation of a situational judgment inventory (SJI) and biographical data (biodata) measure intended to evaluate students' noncognitive attributes and to predict multiple dimensions of college student performance. We also determine the incremental validity of these measures above the validity of the SAT/ACT and existing Big Five personality measures (Digman, 1990). Incremental validity above personality measures is important because personality is a major component of the noncognitive domain, and because there is some evidence that biodata and SJI measures correlate with personality measures (e.g., Clevenger, Pereira, Wiechmann, Schmitt, & Harvey-Schmidt, 2001; Stokes & Cooper, 2001). If existing measures of general personality constructs account for the same variance as newly constructed biodata and SJI measures, then it would be much more economical and theoretically parsimonious to use only personality measures. Additionally, we extend the usual validation study conducted in academic situations by considering not only college GPA but also class attendance and peer and self-ratings across a broad set of performance dimensions reflected in the goal or mission statements of a representative cross-section of American universities. Because these latter outcomes are likely to be more highly related to noncognitive determinants than is GPA, consideration of a broad array of less cognitively loaded predictors should be informative.

Biodata measures provide a structured and systematic method for collecting and scoring information on an individual's background and experience (Mael, 1991; Nickels, 1994). SJIs are multiple-choice tests intended to appraise how an applicant (for a job, or in this case for college) might react in different relevant contexts (Motowidlo & Tippins, 1993). In the college context, both measures have the potential for increased criterion-related validity over traditional measures, because item content can be tailored to specific dimensions of student performance in college and to the goals of a particular college or college-admissions process. Biodata and SJIs may also have greater practical utility over alternative subjective evaluations, such as essays or reference letters that are commonly used in college admissions, because biodata and SJIs provide a fair and standardized method to obtain and score information about the broad range of prior educational and social experiences that applicants may have had. In developng any measure, however, the cost and time liabilities (e.g., developing substantively and psychometrically appropriate items, carefully developing empirical scoring keys) must be weighed against the potential advantages. In this context, however, it is likely the case that objectively scored biodata and SJI instruments will collect such information more efficiently than reading and scoring essays and application blanks.

Expanding the Criterion Space of College Student Performance

Conceptualizing and evaluating the successful development of college students should reflectsome function of the multiple goals and outcomes desired by students, the school administration, legislators, and others with a vested interest (Willingham, 1985). Theoretically, the concern in the educational literature for multiple dimensions of college performance parallels the development of multidimensional models of job performance in the industrial/organizational (I/O) psychology literature (e.g., Borman & Motowidlo, 1997; Campbell et al., 1993). In an early attempt to understand multiple dimensions of college performance systematically, Taber and Hackman (1976) identified 17 academic and nonacademic dimensions to be important in classifying successful and

unsuccessful college students. Examples of these dimensions are intellectual perspective and curiosity, communication proficiency, and ethical behavior. Furthermore, college students actively engaged across numerous domains have tended to achieve greater success in their overall college experience as reflected in their scholastic involvement, accumulated achievement record, and their graduation (Astin, 1984; Willingham, 1985).

Our own effort to identify the number and nature of dimensions of college student performance was an exploratory information-gathering process that followed two primary guidelines. First, the number of dimensions should not be so many that the information is unwieldy, yet not so few that the domain of college performance is oversimplified and not appropriately represented. Second, we wanted to understand how a variety of stakeholders in the process and outcomes of college education define student success in college, because relying on one source alone could lead to biased or deficient definitions and representations of the college performance domain. The 12 dimensions that resulted stem from themes contained within the mission statements and stated educational objectives we sampled across a range of colleges and universities (see the Method section for procedural details). These dimensions are defined in Table 1, and they are referred to in the text in abbreviated form. They deal with intellectual behaviors (Knowledge, Learning, and Artistic), interpersonal behaviors (Multicultural, Leadership, Interpersonal, and Citizenship), and intrapersonal behaviors (Health, Career, Adaptability, Perseverance, and Ethics).

Biodata

Biographical data, or biodata, contain information about one's background and life history (Clifton, Mumford, & Baughman, 1999; Mael, 1991; Nickels, 1994). Despite the informal use of similar information in college applications (e.g., extracurricular

Table 1
Twelve Dimensions of College Performance

Dimension	Definition
	Intellectual behaviors
Knowledge, learning, and mastery of general principles (Knowledge)	Gaining knowledge and mastering facts, ideas, and theories and how they interrelate, and understanding the relevant contexts in which knowledge is developed and applied. Grades or grade point average can indicate, but not guarantee, success on this dimension.
Continuous learning, and intellectual interest and curiosity (Learning)	Being intellectually curious and interested in continuous learning. Actively seeking new ideas and new skills, both in core areas of study and in peripheral or novel areas.
Artistic cultural appreciation and curiosity (Artistic)	Appreciating art and culture, either at an expert level or simply at the level of one who is interested.
	Interpersonal behaviors
Multicultural tolerance and appreciation (Multicultural)	Showing openness, tolerance, and interest in a diversity of individuals (e.g., by culture, ethnicity, or gender). Actively participating in, contributing to, and influencing a multicultural environment.
Leadership (Leadership)	Demonstrating skills in a group, such as motivating others, coordinating groups and tasks, serving as a representative for the group, or otherwise performing a managing role in a group.
Interpersonal skills (Interpersonal)	Communicating and dealing well with others, whether in informal social situations or more formal school-related situations. Being aware of the social dynamics of a situation and responding appropriately.
Social responsibility, citizenship, and involvement (Citizenship)	Being responsible to society and the community and demonstrating good citizenship. Being actively involved in the events in one's surrounding community, which can be at the neighborhood, town/city, state, national, or college/university level. Activities may include volunteer work for the community, attending city council meetings, and voting.
	Intrapersonal behaviors
Physical and psychological health (Health)	Possessing the physical and psychological health required to engage actively in a scholastic environment. This would include participating in healthy behaviors, such as eating properly, exercising regularly, and maintaining healthy personal and academic relations with others, as well as avoiding unhealthy behaviors, such as alcohol/drug abuse, unprotected sex, and ineffective or counterproductive coping behaviors.
Career orientation (Career)	Having a clear sense of career one aspires to enter into, which may happen before entry into college or at any time while in college. Establishing, prioritizing, and following a set of general and specific career-related goals.
Adaptability and life skills (Adaptability)	Adapting to a changing environment (at school or home), dealing well with gradual or sudden and expected or unexpected changes. Being effective in planning one's everyday activities and dealing with novel problems and challenges in life.
Perseverance (Perseverance)	Committing oneself to goals and priorities set, regardless of the difficulties that stand in the way. Goals range from long-term goals (e.g., graduating from college) to short-term goals (e.g., showing up for class every day even when the class is not interesting).
Ethics and integrity (Ethics)	Having a well-developed set of values, and behaving in ways consistent with those values. In everyday life, this probably means being honest, not cheating (on exams or in committed relationships), and having respect for others.

Note. Summary label for each dimension is in parentheses. These labels are used in subsequent tables.

activity lists and resumes), we undertook the development of a biodata inventory with standard multiple-choice responses to questions about one's previous experiences, in a manner similar to that of biodata tests used in employee selection.

Several decades of research in the employment arena have indicated that biodata instruments are usefully related to job performance measures (see Hunter & Hunter, 1984; Mumford & Stokes, 1992; Schmidt & Hunter, 1998; Schmitt, Gooding, Noe, & Kirsch, 1984). Further studies (Brown, 1981; Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990) have explored the degree to which such instruments generalize across companies and industries, and Stokes and Cooper (2001) have also demonstrated that biodata items can be written to reflect meaningful psychological constructs. Perhaps most relevant to the research reported in this article is work reported by Owens and his colleagues (Mumford, Stokes, & Owens, 1990; Owens & Schoenfeldt, 1979; Stokes, Mumford, & Owens, 1989) in which they developed measures of developmental patterns of life history experiences, collected data from large groups of college students, and reported meaningful relationships with a variety of subsequent academic and life outcomes. They also found evidence of considerable stability in these life history patterns over time.

The present study was undertaken for several reasons. First, we conceptualized and modeled the college student performance domain broadly, consistent with the stated objectives of a broad cross-section of U.S. universities. This led to the development of outcome measures corresponding to these performance dimensions, while still including GPA as a traditional measure of student performance. We also used this performance domain as the blueprint by which we developed biodata and situational judgment measures as predictors. Second, we evaluated the psychometric adequacy of these measures. Third, we assessed the relationship between the biodata and SJI measures and college performance outcomes as evidence of their validity. Fourth, in multivariate analyses, we assessed the degree to which the biodata and SJI provided incremental validity over the SAT or ACT and a structured and widely available Big Five personality measure. The latter was important because several of the dimensions measured with the biodata and SJI were similar to personality constructs. Fifth, we used item-level empirical keying methods to develop a subset of biodata and SJI items with high criterion-related validity across samples. Finally, because of the concern with the adverse impact resulting from standardized cognitive ability and achievement tests, we examined mean differences in our instruments across racial and gender subgroups.

Situational Judgment Inventories

Situational judgment inventories (SJIs) are measures in which respondents choose or rate possible actions in response to hypothetical situations or problems. SJIs tend to be less costly to construct and administer than more complex simulations like work samples and assessment centers (Motowidlo, Dunnette, & Carter, 1990). Although SJIs have been in and out of favor in employment contexts for more than 80 years, there has been a renewed interest because of their validity as employment tests designed to predict job performance. A meta-analysis by McDaniel, Bruhn-Finnegan, Morgeson, Campion, and Braverman (2001) estimated that SJIs have an overall criterion-related validity of $\rho = .34$, though there appears to be substantial variability associated with that value ($\sigma_\rho = .14$, with a 90% credibility interval of .09 to .69), with job complexity as a potential moderator (Huffcutt, Weekley, Wiesner, DeGroot, & Jones, 2001). In the employment context, the use of SJIs usually reduces adverse impact for minorities compared with that of cognitive tests (Pulakos & Schmitt, 1996; Sackett et al., 2001), and the SJI produces favorable test-taker reactions (Hedlund et al., 2001) as well as high perceptions of face validity (Clevenger et al., 2001; McDaniel et al., 2001). Such support for SJIs in employment settings suggests that they may be a viable supplement or alternative to traditional cognitive ability testing in college admissions as well, although we are aware of only one previous application in academic prediction. Hedlund et al. (2001) found an SJI to have rather small incremental validity above GMAT scores for MBA students ($\Delta R^2 = .03$). Our SJI development is based on a different set of considerations and methods than was the Hedlund et al. effort, including a broader definition of student performance.

Although the research on SJIs so far indicates that they hold promise as valid predictors of job performance, the construct validity of SJIs remains elusive (Clevenger et al., 2001). Unlike "purer" measures of ability or personality, SJIs reflect complex, contextualized situations and events. It is therefore reasonable to think that constructs measured by SJIs are related yet somewhat different from cognitive ability (Sternberg et al., 2000), having much in common with personality constructs as well, because SJIs rely on individuals' subjective judgments of response-option appropriateness. SJIs are merely measurement methods with content tailored to a particular context, though, and therefore correlations with personality and cognitive ability may vary widely across situations in which SJIs are developed.

Method

Sample

Six hundred fifty-four first-year undergraduate freshman students at a large midwestern university volunteered for this study and received $40 for their participation. Of these, 644 provided usable data after screening for careless responses. Students were recruited through their classes, housing units, and through the student newspaper. Mean age was 18.5 years ($SD = 0.69$). Seventy-two percent were female. Seventy-eight percent were White, 9.5% were African American, 1.9% were Hispanic American, 5.3% were Asian American, and 4.5% were from other racial/ethnic groups. This sample was very nearly identical to the university in terms of racial/ethnic identity: 77.3% were White, 9.8% were African American, 1.9% were Hispanic American, 5.4% were Asian American, and 5.6% other. Our sample overrepresented female students, as 55% of the university's freshman were women. In the admissions process at this university, students completed an application that required the usual admissions materials including the ACT, high school transcript, basic demographic data, previous schools attended, extracurricular activities, and high school honors and activities. Typically neither the educational literature nor colleges themselves indicate clearly how and to what extent information such as activities, awards, and past experiences are used in making actual admissions decisions.

Measures

Dimensions of College Performance

Several of our measures (biodata, SJI, self-rated and peer-rated behaviorally anchored rating scales) were developed on the basis of 12 dimen-

sions of college performance. The process of establishing these dimensions first involved examining the Web pages of colleges and universities, selecting institutions of differing levels of prestige as indicated by *U.S. News and World Report* rankings. Specifically, we read the content of the home pages for 35 colleges and universities, searching for explicitly stated educational objectives or mission statements; if the Web page had a search engine, we also entered relevant search terms. These colleges and universities varied on characteristics such as public/private and large/small enrollment, and 23 institutions provided usable information. Institutions not providing usable information did not explicitly state their educational objectives or provide a university mission statement. There were no apparent systematic differences between those institutions providing usable information and those that did not. The information gathered off the Web pages were parsed into smaller discrete sentence fragments, retaining the original wording as much as possible. For example, the sentence fragment "promote a commitment to learning, freedom, and truth" was decomposed into "promote a commitment to learning," "promote a commitment to freedom," and "promote a commitment to truth." Decomposing these fragments resulted in 174 separate goal statements (including content overlap across institutions). Independently, three of the present authors rationally sorted the statements into as many or as few clusters as they liked; then in a subsequent group meeting, they agreed on 12 dimensions through joint discussion of their independent sorts. It is clear that our sampling from colleges and universities was far from exhaustive; however, it was representative enough so that the college performance domain was truly multidimensional, representing a wide domain of the college experience. It would be difficult to imagine adding many more dimensions to the framework while remaining at this level of construct generality. However, to be sure, we concurrently interviewed a lead administrator at the Michigan State University Department of Residence Life, who provided us with University Residence Life materials that we content analyzed. Finally, criteria identified through our Web search and from university resources were compared against college performance criteria identified in other related educational research (Beatty, Greenwood, & Linn, 1999; Patelis & Camara, 1999; Sackett et al., 2001; Taber & Hackman, 1976; University of Pennsylvania, 2000; Wightman & Jaeger, 1998).

At this point we proceeded with the 12 performance dimensions, and the same three raters then independently re-sorted the goal statements, now reduced to 134 statements because of content redundancies. Of those statements, 85 (62%) were agreed upon by all three raters, and 129 statements (96%) were agreed upon by at least two out of the three raters. After this re-sorting task, each of the identified dimensions was compared with similar dimensions in the industrial and organizational, educational, and vocational psychology literature involving a college population. In some instances, the dimension labels and definitions were modified to be more consistent with the language of the current literature research. The end result of rationally and systematically combining information from these diverse sources resulted in the 12 dimensions of Table 1.

Predictor Measures

Big Five. The Big Five personality traits, also known as the Five-Factor Model (FFM), represent the most commonly, although not universally, accepted personality framework in the current psychological literature (Goldberg, 1993; McCrae & Costa, 1999; Wiggins & Trapnell, 1997). FFM personality traits were measured using a 50-item personality measure from the International Personality Item Pool (Goldberg, 1999). This measure is psychometrically comparable with other commonly used measures of the FFM of personality, such as the NEO–Personality Inventory (Costa & McCrae, 1992). Goldberg (1999) reported the mean coefficient alpha for each of the five scales (10-items each) to be .84, indicating an acceptable degree of internal consistency. Our data were consistent with this, with alphas of .88, .81, .83, .84, and .76, respectively, for the scales of Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (ES, essentially the opposite of Neuroticism), and Openness (O).

Social desirability. The tendency for respondents to give socially desirable responses on noncognitive measures such as the biodata and personality measures is well documented in the social and personality psychology literature (Paulhus, 1988). To assess the degree to which our measures might be susceptible to social desirability, we administered the Paulhus measures of self-deception and impression management. Each measure contained 19 items as we removed 2 items that seemed too intrusive to use in the present context ("I never read sexy books or magazines" and "I have sometimes doubted my ability as a lover"). Paulhus (1991) presented evidence that these measures possess adequate reliability; in our study coefficient alphas were .62 for self-deception and .80 for impression management, indicating marginal and acceptable levels of internal-consistency reliability, respectively.

SAT/ACT. Authorization to obtain SAT or ACT scores was obtained as part of the informed-consent procedures used in the data collection. Out of 644 participants, we obtained 151 SAT scores and 610 ACT scores. All participants had taken one of these tests, and many had taken both as part of their application to different universities. SAT and ACT composite scores were correlated .85; thus, these variables were standardized on national norms within each test and if necessary combined, resulting in a single index of the participants' ability or preparation to do college work.

Biodata. Multiple sources were searched for preexisting biodata items that would relate to the aforementioned 12 performance dimensions (see Table 1). This search identified 197 items whose content was judged to be relevant to one of our dimensions and to the college context. Most of our biodata items were adapted from Pulakos, Schmitt, and Keenan (1994) and Mumford (2001). However, we also reviewed the content of the University of Georgia Biographical Questionnaire (Owens, Albright, & Glennon, 1966), the Assessment of Background and Life Experiences (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990), the Personnel Reaction Blank (Gough & Arvey, 1998), a biographical questionnaire by Russell, Green, and Griggs (n.d.), and a biodata measure developed by Schmitt and Kunce (2002). Items varied in the type of response scale (frequency of behavior, Likert scale) and also in the nature of the constructs measured (past beliefs and attitudes, behaviorally based experiences). All item stems were modified to be appropriate for the college context. After this process, several of our 12 dimensions still lacked a sufficient number of items, so we rationally generated additional items for those dimensions.

Many if not most of the preexisting items had response options that did not apply to the college student population or included inappropriate response options, so item content and response options were rewritten accordingly. Revised items were pilot tested on six paid college students who supplied open-ended responses that were subsequently modified to reflect a reasonable range of response options, dropping items showing little variance or content redundancy with other items.

The stability of the structure of the rational or theoretical inventory was established by assessing interrater agreement on a rational sort of the items. Specifically, six researchers resorted all items back into the 12 dimensions. Items on which five of six raters agreed with the originally assigned dimension were retained; those on which four of six agreed were discussed and rewritten or dropped, and those with less agreement were discarded. When all six raters assigned an item to one dimension other than the one to which it was originally assigned, it was reassigned to that new dimension. Using these criteria, we discarded 5 items and reassigned several to a new dimension. The final biodata inventory then consisted of 115 items representing our 12 dimensions, each scored on a 4- or 5-point scale. Sample biographical data items can be found in Appendix A.

SJI. A search of existing SJI measures led to creating item stems that were adapted to our 12 dimensions of college student performance (see Table 1). We recruited and paid undergraduate students at a large midwestern university to participate in developing our SJI further. First, students generated critical incidents for use as additional item stems for dimensions underrepresented by existing SJI items. Next, an independent set of students created multiple response options for each item stem.

Finally, we developed a scoring key based on responses of advanced (junior and senior) college students in a undergraduate course in psychological measurement, who responded to SJI items as part of a course project ($N = 42$). Each item presents a situation about which students made two judgments indicating which responses would be the "best" and "worst." The scoring key was then developed from these responses. Item response options were part of the scoring key if their means showed statistically significant differences between each other in the frequency endorsed as "best" or "worst" (details of the empirical scoring procedure are in Motowidlo et al., 1990, and Motowidlo, Russell, Carter, & Dunnette, 1988). We then sorted all items for which the scoring key was developed back into our 12 performance dimensions. Items with less than 75% agreement were discarded from the inventory; items were discussed if they had greater than 75% agreement yet were sorted back into a different category. This resulted in a total of 57 items for the final SJI instrument, in which each scale consisted of 3 to 6 items. Individuals could receive a score on each item ranging from +2 (if they agreed with both the "best" and "worst" response keys) to –2 (if they indicated that the "best" item is the worst and the "worst" item is the best). Refer to Appendix B for sample SJI items.

Outcome Measures

GPA. With university authorization, we obtained study participants' GPA from the registrar's office as part of the informed-consent process. First- and second-semester GPAs were obtained for 621 of the respondents; these two values were averaged to yield first-year college GPA measure. First-year GPA was judged to be a useful outcome as part of the domain of college performance, because although it may be less related to long-term outcomes than other criteria, it is a critical criterion for staying in college during the early years (vs. being put on probation or being expelled), though having college-grade data longitudinally would also be of interest in future research.

Absenteeism. Absenteeism was assessed with a single self-report measure whereby participants responded by selecting the approximate number of classes they had missed in the past 6 months on a 5-point scale ranging from *less than 5* to *more than 30.*

Behaviorally anchored rating scale for multiple dimensions of college performance. The 12 dimensions served as a guide in developing a behaviorally anchored rating scale (BARS). For each of the 12 BARS items, a dimension name and its definition were presented along with two examples of college-related critical incidents and various behavioral anchors that reflected three levels of performance on a 7-point scale, which ranged from *unsatisfactory* to *exceptional.* Both critical incidents and anchors were taken from the incidents and response options generated during SJI development. Also, we collected data on a peer- and self-rated version of these BARS, both of which referred to a student's performance in college. See Appendix C for sample BARS items.

Self-rated BARS items were administered after the biodata and SJI questions in a larger test administration described below. Dimensionality in these ratings would provide some evidence for distinct dimensions of performance, as was found in past studies of task and contextual dimensions of job performance (LePine, Erez, & Johnson, 2002; Rotundo & Sackett, 2002). However, a principal-axis exploratory factor analysis (EFA) yielded a large first factor that accounted for 32% of the variance and four times as much variance as the second factor. Multiple-factor models did not provide a readily interpretable solution. A confirmatory factor analysis (CFA) of these ratings using LISREL 8.51 (Jöreskog & Sörbom, 2001) yielded support for a single-factor model, $\chi^2(54, N = 641) = 122.71$, $p < .01$; root-mean-square error of approximation (RMSEA) = .05; comparative fit index (CFI) = .95; and nonnormed fit index (NNFI) = .93. Coefficient alpha for the 12 BARS ratings was .80. Thus, subsequent data analyses used a composite rating based on the mean of the 12 BARS items.

The 12 peer-rated BARS items were identical to the self-rated BARS except for appropriate wording changes from first to third person. During the test administration, study participants were asked to nominate a peer who knew their work well and could provide ratings of the participant on the same 12 dimensions. Follow-up contacts of these peers by e-mail led to ratings of 145 participants, and raters were compensated $5 for their participation. Note that most of the peers were friends or roommates (83%); other categories of peers included resident assistants, teaching assistants, or professors. Peers were largely known after 6 months to a year of college (40%) or they were known for 3 years or more (40%). Subgroups, both by peer-rater type and by length of acquaintanceship, were too small to allow for meaningful post hoc statistical tests (e.g., for differences in correlations); however certainly different types of peer raters may provide different sources of insight into the student being rated. Similar to the self ratings, EFA and CFA analyses of these peer ratings using LISREL 8.51 yielded support for a single-factor model, $\chi^2(54, N = 145) = 84.38$, $p < .01$, RMSEA = .06, CFI = .93, NNFI = .91, and coefficient alpha of the composite of the peer ratings was .83. The fourth measure of student performance was this composite peer rating, which also was an average across BARS dimension ratings; the composite peer rating was available only on these 145 respondents.

Administration of the Paper-and-Pencil Tests

All of the measures were administered with a series of four booklets, in small group administrations ($M = 15.19$ participants, $SD = 8.12$). Participants were provided with test booklets and machine-scannable answer sheets. Trained proctors adhered to a script and read test instructions verbatim, similar to standardized test procedures. Sessions were scheduled to last 4 hr, allowing participants sufficient time to complete the various measures. Breaks were held after the administration of the first and second booklets.

Two forms of the test were randomly assigned, Form A and Form B. The two forms were identical, apart from the requirement that some of the biodata questions on Form B required that participants elaborate in writing in support of their multiple-choice response. Written responses were not scored; they were requested as part of an effort to control for the impact of social desirability; the results of this effort are reported in another study (Schmitt et al., 2003). Random assignment of test forms was done by group so that test-taking experience would be similar within each group, and participants would not notice some people were doing substantially more or less writing than others. Written elaboration did have an impact on the descriptive statistics for some of the biodata responses, so it was appropriate to standardize all items within form before creating composites or conducting correlational analyses.

Results

Scale-level descriptive statistics, reliability coefficients, and intercorrelations for our 12 college performance dimensions are shown for the biodata and SJI scales in Tables 2 and 3, respectively. For the biodata scales, most coefficient alphas are acceptable (above .70) or marginal (Perseverance $\alpha = .63$), and although the Interpersonal, Career, and Adaptability alpha reliabilities are poor, the latter two scales had fewer than the usual 10 items because of subsequent item and scale refinement. Intercorrelations between biodata scales did not approach the reliability of the scales in almost all cases, indicating reasonable levels of discriminant validity (see Table 2, above the diagonal). Correlations with self-deception and impression management are modest in most cases, though the correlation between the Ethics scale and impression management is quite high ($r = –.54$). In general, these correlations indicate that the two social desirability dimensions covary with

Table 2
Biodata Scales: Descriptive Statistics, Intercorrelations Between Scales and Correlations With Self-Deception and Impression Management

Scale	*k*	*SD*[a]	1	2	3	4	5	6	7	8	9	10	11	12	S-D	IM
1. Knowledge	10	5.31	*.72*	.75	.47	.56	.43	.42	.51	.40	.37	.29	.80	.47	.29	−.28
2. Learning	9	4.72	.52	*.67*	.78	.88	.56	.38	.63	.26	.22	.40	.35	.25	.29	−.18
3. Artistic	9	5.97	.37	.58	*.84*	.89	.41	.24	.57	.03	.09	.15	.23	.09	.18	−.11
4. Multicultural	10	5.66	.41	.63	.71	*.76*	.56	.31	.63	.13	.12	.29	.37	.13	.16	−.12
5. Leadership	10	5.92	.32	.41	.33	.44	*.79*	.50	.76	.39	.23	.33	.58	.10	.18	−.11
6. Interpersonal	10	4.16	.25	.21	.15	.19	.30	*.47*	.32	.57	.22	.76	.57	.30	.32	−.26
7. Citizenship	10	5.26	.37	.44	.44	.46	.57	.19	*.71*	.21	.40	.15	.53	.30	.19	−.22
8. Health	10	5.25	.29	.18	.02	.10	.29	.33	.15	*.71*	.13	.60	.57	.21	.25	−.16
9. Career	3	2.15	.23	.13	.06	.07	.15	.11	.24	.08	*.53*	.11	.58	.35	.21	−.15
10. Adaptability	7	3.74	.32	.25	.11	.19	.22	.40	.09	.39	.06	*.58*	.60	.24	.36	−.18
11. Perseverance	8	4.22	.54	.36	.17	.25	.41	.31	.36	.38	.34	.36	*.63*	.48	.35	−.29
12. Ethics	6	3.86	.34	.17	.09	.10	.08	.18	.22	.15	.22	.15	.32	*.72*	.21	−.54

Note. $N = 638$. Coefficient alpha reliability coefficients are italicized on the main diagonal; observed correlations are in the lower triangle, and correlations corrected for attenuation due to measurement unreliability are in the upper triangle. Refer to Table 1 for definitions of scales. *k* = number of items; S-D = self-deception; IM = impression management.
[a] Means of the biodata scales are all near zero because Forms A and B were standardized before computing composites; Form B had some items requiring written elaboration to multiple-choice responses.

these biodata measures. This covariation may be a problem for the use of biodata measures as the basis for admission decisions. However, recent assessments of the correlation of social desirability (Ones, Viswesvaran, & Reiss, 1996) and impression management (Viswesvaran, Ones, & Hough, 2001) with job performance indicate these correlations are near zero. Also across four employment data sets, Ellingson, Smith, and Sackett (2001) found consistent and fairly large mean differences on personality measures between high scorers and low scorers on a scale of socially desirable responding, though the factor structure between personality predictors remained stable. The implication of these studies is that, at least in the employment context, social desirability may correlate with various predictors and influence their means in a socially desirable direction, but neither the correlation nor the mean increase appears to impact criterion-related validity greatly. Other strategies for detecting faking or socially desirable responding (Stark, Chernyshenko, Chan, & Lee, 2001) may have different implications for criterion-related validity.

In contrast with Table 2, the data in Table 3 regarding the psychometric adequacy of the SJI are much less encouraging. The coefficient alphas are low, and intercorrelations between the scales indicate a lack of evidence for discriminant validity. This may be a function of the small number of items in each scale and the general complexity of situational-judgment items, and it is also consistent with previous research on situational judgment measures that usually treats items together as a single construct (McDaniel et al., 2001). Correlations between the SJI scales with self-deception are low relative to the biodata scales, and for Ethics the SJI scale correlates with impression management lower than the biodata scale does, but these findings may be partly a function

Table 3
SJI Scales: Descriptive Statistics, Intercorrelations Between Scales, and Correlations With Self-Deception and Impression Management

Scale	*k*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	S-D	IM
1. Knowledge	3	2.726	2.01	*.37*	.79	.73	.53	.69	.60	.65	1.02	.72	.38	1.05	.60	.14	−.24
2. Learning	5	2.849	2.50	.23	*.23*	.98	.99	.78	1.00	.76	1.59	.95	1.14	.26	.85	.08	−.13
3. Artistic	5	1.981	2.92	.28	.29	*.39*	1.04	.84	.83	.96	.96	.58	.97	.70	.21	.17	−.14
4. Multicultural	5	3.609	2.69	.20	.30	.41	*.39*	.71	.90	.84	.97	.67	.76	.55	.58	.13	−.19
5. Leadership	5	3.244	3.00	.28	.25	.35	.29	*.44*	.94	.83	.99	.60	.90	.78	.54	.19	−.16
6. Interpersonal	4	2.273	2.29	.21	.28	.30	.33	.36	*.34*	1.00	1.10	.65	.87	.70	.71	.09	−.21
7. Citizenship	5	2.423	2.53	.22	.21	.34	.30	.31	.33	*.32*	1.14	.45	.75	.70	.68	.16	−.24
8. Health	4	2.845	2.15	.29	.36	.28	.28	.31	.30	.30	*.22*	1.06	1.37	1.17	.86	.06	−.22
9. Career	5	4.978	2.71	.28	.29	.23	.27	.25	.24	.16	.31	*.40*	.95	.81	.54	.03	−.10
10. Adaptability	5	5.207	2.86	.40	.31	.35	.27	.34	.29	.24	.37	.35	*.33*	1.05	.66	.14	−.18
11. Perseverance	5	1.495	2.84	.41	.19	.28	.22	.34	.26	.26	.36	.33	.39	*.42*	.48	.15	−.22
12. Ethics	6	3.622	3.15	.27	.30	.18	.27	.26	.31	.29	.30	.25	.28	.23	*.55*	.07	−.30

Note. $N = 634–642$. Coefficient alpha reliability coefficients are italicized on the main diagonal; observed correlations are in the lower triangle, and correlations corrected for attenuation due to measurement unreliability are in the upper triangle. Corrected correlations are point estimates, and some exceed 1.0. Refer to Table 1 for definitions of scales. SJI = situational judgment inventory. *k* = number of items; S-D = self-deception; IM = impression management.

of the lower internal-consistency reliability of the SJI measures, which would systematically attenuate observed correlations. An EFA of the SJI items revealed the presence of a relatively large general factor accounting for three times the variance of the second factor. Additional factors accounted for small portions of variance, however, and also were difficult to interpret substantively. Because of the lack of discriminant validity and internal consistency of the SJI subscales, we computed a composite SJI index. This measure had high internal consistency reliability ($\alpha = .85$), suggesting that although it was clearly appropriate to sample content representatively across the 12 dimensions of college performance, the corresponding SJI scales are best used as an overall composite reflecting judgment across a variety of situations relevant to college life. The nature of the constructs being measured with this SJI composite can be understood by examining its correlates in the tables. Table 4 also presents the observed correlations between the SJI composite with the biodata scales, showing that the SJI and biodata are correlated yet are distinct, with each having the potential for incremental validity in the prediction of student performance outcomes.

Correlations of Biodata and SJI With SAT/ACT and Big Five Measures

Table 5 reports correlations between the SJI composite and biodata with the Big Five (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness) and the SAT/ACT measure. These correlations are important for two reasons. First, they provide some evidence for the construct validity of the biodata and SJI measures (i.e., the scales are part of a theoretically sensible nomological net). Second, if these correlations are too high, these measures provide redundant information already available through standardized instruments. Correlations between personality measures and the biodata measures were, in fact, relatively high, yet not so high as to preclude the possibility that biodata and the SJI composite will add incrementally to the prediction of student performance outcomes. The strongest positive correlations were found between Extraversion with biodata Leadership ($r = .48$) and biodata Interpersonal ($r = .45$); Agreeableness with biodata Interpersonal ($r = .33$); Conscientiousness with biodata Knowledge ($r = .40$), biodata Perseverance ($r = .47$), and biodata Ethics ($r = .30$); Emotional Stability with biodata Interpersonal ($r = .36$), biodata Health ($r = .33$) and biodata Adaptability ($r = .35$); and Openness with biodata Knowledge ($r = .34$), biodata Learning ($r = .52$), and biodata Artistic ($r = .50$). None of the correlations involving the SAT/ACT measure were higher than .12. In general, the magnitude of the correlations is consistent with the constructs being correlated.

Table 5
Biodata and SJI: Correlations With the IPIP (Big Five) and SAT/ACT

Scale	E	A	C	ES	O	SAT/ACT
Biodata						
Knowledge	.14	.26	.40	.13	.34	.06
Learning	.20	.21	.15	.18	.52	.07
Artistic	.19	.23	.02	.11	.50	.07
Multicultural	.23	.23	.03	.12	.38	.10
Leadership	.48	.23	.15	.10	.24	.05
Interpersonal	.45	.33	.16	.36	.16	−.06
Citizenship	.22	.28	.16	.11	.26	−.01
Health	.17	.09	.26	.33	.06	.01
Career	.06	.13	.23	−.01	.08	−.12
Adaptability	.26	.17	.27	.35	.20	.05
Perseverance	.24	.27	.47	.16	.23	−.10
Ethics	−.12	.25	.30	.12	.11	.00
SJI						
SJI composite	.17	.38	.28	.17	.21	−.03

Note. All correlations with magnitudes above .06 are statistically significant at $p < .05$; all correlations with magnitudes above .09 are statistically significant at $p < .01$. Refer to Table 1 for definitions of scales. SJI = situational judgment inventory; IPIP = International Personality Item Pool; E = Extraversion; A = Agreeableness; C = Conscientiousness; ES = Emotional Stability (Neuroticism); O = Openness to Experience.

Table 4
Correlations Between Biodata and SJI Scales

	Biodata											
SJI	1	2	3	4	5	6	7	8	9	10	11	12
1. Knowledge	.27	.11	.03	.09	.10	.10	.16	.20	.19	.21	.31	.30
2. Learning	.25	.20	.16	.20	.17	.18	.16	.10	.07	.14	.19	.20
3. Artistic	.28	.31	.35	.33	.14	.17	.24	.07	.13	.20	.20	.18
4. Multicultural	.25	.30	.40	.35	.22	.20	.32	.08	.12	.12	.19	.26
5. Leadership	.24	.21	.12	.17	.26	.24	.24	.14	.15	.20	.28	.22
6. Interpersonal	.19	.10	.11	.19	.17	.17	.22	.08	.17	.16	.18	.26
7. Citizenship	.27	.18	.14	.21	.18	.18	.28	.10	.16	.10	.22	.25
8. Health	.24	.13	.07	.10	.09	.17	.13	.22	.13	.11	.26	.22
9. Career	.16	.13	.13	.14	.09	.13	.11	.09	.10	.15	.17	.19
10. Adaptability	.23	.17	.12	.13	.11	.14	.12	.14	.16	.16	.25	.25
11. Perseverance	.24	.11	.01	.10	.19	.15	.20	.13	.13	.15	.31	.24
12. Ethics	.28	.18	.13	.19	.17	.13	.23	.15	.06	.14	.21	.44
SJI composite	.41	.30	.25	.31	.27	.28	.34	.21	.23	.27	.39	.43

Note. $N = 635$. All correlations with magnitudes above .06 are statistically significant at $p < .05$; all correlations with magnitudes above .09 are statistically significant at $p < .01$. Refer to Table 1 for definitions of scales. SJI = situational judgment inventory.

Correlations With Outcome Measures

Table 6 presents correlations between the outcome measures, which reveal that all four outcomes provide distinct but related information regarding student performance. The highest correlation is –.53, the correlation between class absences and first-year GPA. The table also correlates the biodata scales, composite SJI measure, SAT/ACT, and Big Five with the primary outcome measures in our study. These outcomes include the first-year GPA as reported by the university registrar's office, a self-reported index of absenteeism from class, composite self-ratings of performance on the BARS measure, and composite peer ratings on the same BARS measure.

Several biodata scales, such as Knowledge, Health, and Adaptability, do correlate reasonably well with GPA. Several, including the Knowledge, Health, Adaptability, Perseverance, and Ethics scales, also predict class absences (r = –.15 to –.31). Their best correlations, however, are with the composite self-ratings. These higher correlations are probably due to a combination of two factors. First, the biodata measures were designed to reflect the same dimensions that were rated. Second, both the BARS and the biodata measures were completed by the participants. This is not true of the BARS peer ratings on the same dimensions, and correlations between the biodata scales and corresponding BARS peer ratings are lower than similar correlations with BARS self-ratings. Even so, several of the validity coefficients with the peer-rating BARS composite are above r = .20 (i.e., Knowledge, Leadership, Interpersonal, Citizenship, and Perseverance). As for the SJI composite, it is significantly and relatively highly correlated with the self-rating measures of student performance and absenteeism; correlations with GPA and the peer-rating BARS were relatively low. The SAT/ACT measure was correlated with GPA (r = .33), comparable with the uncorrected validity usually displayed in the research summarized in the introduction. Conscientiousness was the only Big Five measure that demonstrated consistent criterion-related validity with correlations ranging from .21 to .30 in absolute magnitude. All five personality scales were significantly related ($p < .05$) to the self-ratings measure, but again, these validities are likely inflated to some degree because both the predictors and the ratings come from the same source.

Table 6
College Performance Outcomes: Intercorrelations and Correlations With Predictors

Variable	*M*	*SD*	GPA	Absenteeism	Self-rating BARS	Peer-rating BARS
			Intercorrelations[a]			
GPA	3.02	0.69	—			
Absenteeism	1.98	1.08	−.53	—		
Self-rating BARS	4.88	0.71	.22	−.22	—	
Peer-rating BARS	4.96	0.80	.29	−.16	−.10	—
			Correlations between predictors and outcomes[b]			
Biodata						
Knowledge			.22	−.19	.47	.21
Learning			.06	.00	.40	.06
Artistic			.01	.07	.37	.09
Multicultural			.08	.01	.38	.10
Leadership			.14	−.04	.41	.20
Interpersonal			.04	−.09	.25	.21
Citizenship			.08	−.08	.39	.22
Health			.24	−.23	.22	.11
Career			−.02	−.07	.17	−.01
Adaptability			.21	−.15	.24	.13
Perseverance			.16	−.21	.45	.21
Ethics			.14	−.31	.35	.02
SJI composite			.16	−.27	.53	.16
SAT/ACT			.33	.11	−.01	.09
Big Five						
Extraversion			−.03	.10	.24	.12
Agreeableness			.10	−.05	.37	.06
Conscientiousness			.21	−.27	.30	.22
Emotional stability			.07	−.05	.15	−.08
Openness			.03	.04	.35	.13

Note. GPA = grade point average; BARS = behaviorally anchored rating scale; SJI = situational judgment inventory.
[a] N = 136. $|r| > .17$ are statistically significant ($p < .05$).
[b] N = 611–636 for correlations with the first three criteria (GPA, absenteeism, and self-rating BARS), where $|r| > .08$ is statistically significant ($p < .05$). N = 136 for correlations with the peer-rating BARS composite, where $|r| > .16$ is statistically significant ($p < .05$).

Development and Double Cross-Validation of Empirically Keyed Items

Identical empirical-keying procedures were carried out on both the SJI and the biodata, each at the item level. First, all cases were randomly split into two samples, in which the *developmental* sample had an *N* of 314 and the *holdout* sample had an *N* of 330 (these sample labels are arbitrary). Second, using the cases within each subgroup, all biodata and SJI items were correlated with three criteria: first-year GPA, absenteeism, and the summed composite of the self-report BARS. The peer-rating composite was not used in these analyses because the sample size was considered too small to allow for meaningful cross-validation (i.e., $N = 147$ for all items). Third, applying a minimum cutoff value to the distribution of these item–criterion correlations produced a single set of "empirically best" biodata and SJI items for each of the three criteria, which resulted in three sets of items for each sample. In all these instances, the cutoff values resulted in the selection of 20 to 40 items.

The sample data used to derive empirical keys may capitalize on chance if they were then correlated with criteria. To avoid this possibility, we applied cross-validation procedures to our data. For cross-validation of the keys, each of the GPA-, absenteeism-, and BARS-based biodata and SJI item sets derived from the developmental sample were then used in the holdout sample to form similar item composite scores, which were then correlated with GPA, absenteeism, and composite self-rating criteria. Note that these holdout validity estimates are shrunken estimates; they are lower because the data are from the holdout sample, and the items were empirically selected based on data in the development sample. This cross-validation procedure was also applied in reverse, resulting in two cross-validated estimates for each of the criteria, one based on the best items chosen from the developmental sample and one based on items from the holdout sample. The two validities were then averaged to provide a single cross-validated estimate for each criterion.

Procedures for selecting the empirically best items for each criterion are separate from those described in the double cross-validation, though they build from the sample-specific item–criterion correlations derived as part of the double cross-validation effort. The final sets of empirically keyed items for each criterion were chosen on the basis of the compound probabilities associated with the Pearson correlation of each item for both the developmental and the holdout samples. For each item, item-criterion *p* values in the developmental sample were combined with *p* values in the holdout sample using the compound probability formula provided by Guilford (1954, p. 440). The formula produced chi-square values representative of the probability that the item–total correlation could occur by chance in both samples (a high chi-square indicates a low probability). Because a large number of items had compound probabilities that were highly significant, cutoffs were made more stringent so that a manageable number of best items would be retained. For biodata items keyed to GPA, the cutoff was a chi-square associated with $p < .001$.

Table 7 provides the average validities for each of the empirically derived scales for each of our three major outcomes. As can be seen, the validities were quite respectable, rivaling those for high school GPA and SAT/ACT scores predicting the first-year GPA outcome. Biodata and SJI scales were also highly correlated with absenteeism. Finally, as expected, the best correlations were with the self-rating composite. Given that we used double cross-validation to develop these scales, it should be emphasized that these validity estimates do not capitalize on sampling error variance and are likely representative of what one would achieve with similar samples from the same population of students.

Table 7
Criterion-Related Validities of Empirically Keyed Scales

	Outcome measure		
Variable	GPA	Absenteeism	Self-rating BARS (total)
SJI keys			
GPA	.23	−.31	.50
Absenteeism	.20	−.33	.47
BARS	.14	−.20	.51
Biodata keys			
GPA	.37	−.27	.47
Absenteeism	.26	−.30	.50
BARS	.15	−.13	.57

Note. $N = 302–328$. The validities are averaged double cross-validities as described in the text. All correlations are significant at $p < .05$. GPA = grade point average; BARS = behaviorally anchored rating scale; SJI = situational judgment inventory.

Multivariate Analyses

To this point, we have provided descriptive data and correlations describing the relationship between the experimental biodata and SJI measures, outcome variables, and standard measures of scholastic competence (SAT/ACT) and personality. This section describes multivariate analyses of the relationships between these variables. These analyses provide information as to which of the new measures have the most criterion-related validity and how they work in combination with traditional predictors of college student performance.

A series of hierarchical regressions tested the incremental validity of our measures using the biodata scales and the SJI composite. We did not use the best empirical composites of the biodata and SJI items developed and described in the previous section, because these composites were developed based on the same sample used here, and therefore their incremental relationship with the four outcomes would capitalize on chance. In these regressions, SAT/ACT and personality were entered in Steps 1 and 2 respectively; in Step 3, the biodata and SJI measures were entered. The same stepwise regressions were used for each of the four outcomes. The results of these regression analyses are presented in Table 8 including the R^2 for each step, and $\triangle R^2$ for Steps 2 and 3.

As expected, the SAT/ACT score predicted first-year GPA; it was also positively and negatively related to absenteeism but substantially less so; however, it did not significantly predict either set of ratings. The Big Five measures added significantly to the prediction of all four outcomes including GPA. The largest increment to prediction was observed for the self-ratings measure, part of which may be due to the fact that both sets of measures came from the same source. Of the personality scales, Conscientiousness was the most consistent predictor, where only in the case of the self-rating was the regression weight nonsignificant. Because ab-

Table 8
Incremental Validity of Biodata and SJI Measures: Hierarchical Regression Results

	GPA		Absenteeism		Self-rating BARS		Peer-rating BARS	
Step and measure	$\hat{\beta}_i$	ΔR^2	$\hat{\beta}_i$	ΔR^2	$\hat{\beta}_i$	ΔR^2	$\hat{\beta}_i$	ΔR^2
Step 1								
SAT/ACT	.344*	.105*	.086*	.012*	−.039	.006	.066	.004
Step 2								
Extraversion	−.132*		.099*		.027		−.017	
Agreeableness	.063		.044		.083*		−.140	
Conscientiousness	.152*		−.139*		.042		.254*	
Emotional stability	−.015		.040		.004		.200*	
Openness	−.089	.080*	−.005	.092*	.092*	.233*	.133	.094*
Step 3								
Knowledge	.091		−.048		.086*		.067	
Learning	−.089		.095		.008		−.273*	
Artistic	−.021		.088		.108*		.084	
Multicultural	.033		−.012		.023		−.001	
Leadership	.111*		.019		.170*		.104	
Interpersonal	−.045		−.006		−.035		.287*	
Citizenship	−.001		−.052		−.024		.236*	
Health	.151*		−.167*		.008		.077	
Career	−.063		.060		−.020		−.104	
Adaptability	.105*		−.047		−.017		−.046	
Perseverance	−.014		.016		.129*		.048	
Ethics	.018		−.192*		.113*		−.054	
SJI	.047	.062*	.164*	.193*	.260*	.216*	−.084	.136*
R		.50		.47		.67		.48
Adjusted *R*		.47		.44		.66		.34
N		609		627		625		144

Note. $N = 611–636$ for correlations with the GPA, absenteeism, and self-rated BARS criteria; $N = 136$ for correlations with the peer-rating BARS composite. Refer to Table 1 for definitions of scales. SJI = situational judgment inventory; GPA = grade point average; BARS = behaviorally anchored rating scale.
* $p < .05$.

senteeism is a measure of the number of classes missed, the negative regression weight for that outcome means that those high on Conscientiousness tended to miss fewer classes. The regression weight for Extraversion was significant in the prediction of both GPA (a negative relationship) and absenteeism (a positive relationship). Regression weights for Agreeableness and Emotional Stability were significant for self-ratings and peer ratings, respectively.

The biodata measures and the SJI composite added a statistically significant increment to the prediction of all four outcomes. Even for predicting GPA, the R^2 increment was relatively large ($\Delta R^2 = .062$). The SJI regression weight was statistically significant in the absenteeism and self-rating equations. Leadership, Health, and Adaptability regression weights were significant in the GPA equation, and Health and Ethics were significant in the absenteeism equation. Tests for incremental prediction at this step in the hierarchical regressions are particularly conservative because the equations already include the major cognitive ability predictor and the five personality measures. Several biodata scales (Knowledge, Artistic Appreciation, Leadership, Perseverance, and Ethics) were related to the self-ratings. As mentioned, the latter may be somewhat inflated due to the fact that both sets of measures were completed by the students, but when peer ratings were the outcome variable, we also found several significant biodata predictors (Interpersonal, Citizenship, and Learning). Results for the peer-rating criterion were based on a substantially smaller sample ($N = 144$).

Adjusted multiple *R*s for all four equations were substantial ($R = .47, .44, .66$, and .34) and practically significant by most standards. Incremental changes in squared multiple correlations were also fairly large for all four outcomes ($\Delta R^2 = .062, .193, .216$, and .136), indicating the potential utility of the biodata and SJI as novel predictors of various measures of college student performance.

Subgroup Mean Differences

One motivation for searching for predictors of academic performance other than SAT/ACT is that there are large subgroup differences in SAT/ACT scores. These differences often mean that members of minority groups are denied admission to college at greater rates than are members of majority groups. For the SAT/ACT, these mean differences are largest for ethnic/racial subgroups; for noncognitive measures, the research literature does find moderate male/female mean differences, depending on the construct measured (Hough et al., 2001). Hence, one important factor in considering the use of the biodata and SJI measures as alternative predictors of performance in college is the magnitude of any subgroup mean differences.

Table 9 presents male and female means for biodata and for the SJI composite. Significant mean differences in biodata scores were found on almost all scales. Female students tended to score higher across most biodata scales: Knowledge, Leadership, Interpersonal,

Table 9
Biodata and SJI Scores: Descriptive Statistics and Mean Gender Differences

Scale and gender	*n*	*M*	*SD*	*d*
Biodata				
Knowledge				.28*
Male	177	−.11	.54	
Female	461	.04	.52	
Total	638	.00	.53	
Learning				−.18*
Male	177	.07	.58	
Female	461	−.03	.50	
Total	638	.00	.52	
Artistic				.23*
Male	177	−.11	.70	
Female	461	.04	.65	
Total	638	.00	.67	
Multicultural				.13
Male	177	−.05	.53	
Female	461	.02	.57	
Total	638	.00	.56	
Leadership				.31*
Male	177	−.13	.54	
Female	461	.05	.60	
Total	638	.00	.59	
Interpersonal				.41*
Male	177	−.12	.41	
Female	461	.04	.41	
Total	638	.00	.42	
Citizenship				.30*
Male	177	−.11	.50	
Female	461	.04	.53	
Total	638	.00	.53	
Health				−.19*
Male	177	.07	.57	
Female	461	−.03	.51	
Total	638	.00	.52	
Career				.51*
Male	177	−.26	.71	
Female	461	.10	.70	
Total	638	.00	.72	
Adaptability				.09
Male	177	−.04	.54	
Female	461	.01	.53	
Total	638	.00	.53	
Perseverance				.31*
Male	177	−.12	.54	
Female	461	.04	.52	
Total	638	.00	.53	
Ethics				.44*
Male	177	−.20	.69	
Female	461	.08	.61	
Total	638	.00	.64	
SJI composite				.70*
Male	177	.50	.35	
Female	463	.72	.30	
Total	640	.66	.33	

Note. Refer to Table 1 for definitions of scales. SJI = situational judgment inventory. Positive values of *d* favor females, and negative values of *d* favor males.
[a] Overall means of the biodata scales were all near zero because responses were standardized before computing composites to remove the effects of elaboration.
* $p < .05$.

Citizenship, Career, Perseverance, and Ethics had *d* values with magnitudes at or above .30. For the SJI composite, the effect size for the female mean relative to the male mean was also rather high ($d = .70$); female students scored slightly lower than male students on the biodata scales of Learning ($d = –.18$) and Health ($d = –.19$). Results for mean gender differences on the personality measures of the Big Five show that female students tended to score significantly higher than male students on Extraversion ($d = .26$), Agreeableness ($d = .57$), and Conscientiousness ($d = .30$), and lower than male students on Openness ($d = –.27$) and Emotional Stability ($d = –.18$). Effect sizes are moderate with the exception of that for Agreeableness ($d = .57$), indicating that female students tend to score much higher than male students on this scale. In contrast with these differences on noncognitive measures, female students tended to score lower than male students ($d = –.29$) on the SAT/ACT variable. Despite this slightly lower mean on the cognitive ability measure, however, women tended to outperform men on all four major college performance outcomes: GPA ($d = .11$), absenteeism ($d = .25$), self-rating BARS ($d = .39$), and peer-rating BARS ($d = .19$).

Table 10 presents descriptive statistics and standardized mean differences on the biodata scales and SJI composite for four major racial subgroups (White, African American, Hispanic, and Asian American). Before discussing these differences, we caution that the numbers in the minority groups are all relatively small. Analysis of variance by racial group showed significant mean differences in some areas. On biodata scales, Blacks showed a lower mean score than Whites for Health, and a higher score for Career. For most biodata scales, subgroup differences are relatively small (absolute values of $d < .20$) with the exception of the higher mean score for Hispanics on the Learning scale ($d = .63$). On the SJI composite, all three minority subgroups had mean values that were comparable with Whites. As for measures of the Big Five personality scales, the only significant difference between racial subgroups indicated that African Americans tended to score lower than Whites on Agreeableness ($d = –.31$); Hispanics had higher means than Whites on Emotional Stability ($d = .41$) as well as Openness ($d = .44$); and Asian Americans tended to score moderately lower than Whites ($d = –.20$ to $–.25$) across all Big Five scales. Again, subgroup sample sizes were small, and analysis of variance (ANOVA) comparing Whites and minority groups revealed statistically significant mean differences on only the Agreeableness factor. In contrast with the noncognitive measures, mean differences on the SAT/ACT variable were larger in magnitude, as expected from estimates in prior literature. The Black–White standardized mean difference in SAT/ACT scores was $d = –1.22$; the Hispanic–White difference was $d = −1.14$; and the Asian American–White difference was $d = –.36$. In all these cases, the mean scores of White students were higher.

There were also relatively large racial subgroup mean differences on the outcome measures. For the first-year GPA, African American means were lower than Whites ($d = −1.09$), as were Asian Americans ($d = −1.01$) and Hispanics ($d = −1.07$). For the absenteeism measure, comparisons with Whites were not statistically significant for African Americans ($d = –.09$) and Hispanics ($d = .07$), but Asian American students tended to report significantly more absences ($d = .86$). The self-rating BARS performance measure did provide some trends: White and African American means were almost identical ($d = –.03$), but Asian Americans tended to rate themselves lower ($d = –.48$) and Hispanics slightly higher ($d = .15$) than Whites. For the peer-rating BARS performance measure, the numbers of minority subgroup students were too small to provide any statistically significant comparisons. As

Table 10
Biodata and SJI: Descriptive Statistics and Mean Racial Subgroup Differences

Scale	*N*	*M*[a]	*SD*	*d*
Biodata				
Knowledge				
White	501	.01	.52	
Black	60	−.03	.50	−.08
Hispanic	15	−.10	.71	−.20
Asian	34	−.12	.62	−.25
Total	610	−.01	.53	
Learning				
White	501	−.01	.51	
Black	60	.00	.51	.01
Hispanic	15	.32	.62	.63*
Asian	34	−.10	.62	−.19
Total	610	.00	.52	
Artistic				
White	501	−.01	.67	
Black	60	−.14	.63	−.19
Hispanic	15	.47	.60	.73*
Asian	34	.08	.60	.15
Total	610	−.01	.66	
Multicultural				
White	501	−.01	.56	
Black	60	−.07	.53	−.11
Hispanic	15	.34	.55	.63*
Asian	34	.00	.52	.02
Total	610	−.01	.56	
Leadership				
White	501	.02	.58	
Black	60	−.08	.57	−.18
Hispanic	15	.07	.69	.08
Asian	34	−.15	.60	−.30
Total	610	.00	.58	
Interpersonal				
White	501	.02	.41	
Black	60	−.06	.41	−.18
Hispanic	15	.15	.40	.33
Asian	34	−.14	.45	−.38*
Total	610	.01	.41	
SJI composite				
White	504	.67	.32	
Black	60	.68	.35	−.05
Hispanic	15	.71	.35	−.14
Asian	34	.60	.35	−.21
Total	613	.67	.33	

Scale	*N*	*M*	*SD*	*d*
Citizenship				
White	501	.00	.52	
Black	60	.02	.51	.05
Hispanic	15	.12	.66	.23
Asian	34	−.08	.57	−.14
Total	610	.00	.53	
Health				
White	501	.04	.52	
Black	60	−.12	.47	−.31*
Hispanic	15	.07	.56	.06
Asian	34	−.31	.52	−.67*
Total	610	.00	.52	
Career				
White	501	−.04	.72	
Black	60	.20	.68	.34*
Hispanic	15	.36	.63	.56*
Asian	34	.06	.79	.14
Total	610	.00	.72	
Adaptability				
White	501	.01	.52	
Black	60	.03	.52	.03
Hispanic	15	.06	.74	.09
Asian	34	−.20	.57	−.41*
Total	610	.00	.53	
Perseverance				
White	501	−.01	.52	
Black	60	.06	.53	.13
Hispanic	15	.28	.59	.55*
Asian	34	−.11	.57	−.18
Total	610	.00	.53	
Ethics				
White	501	.00	.64	
Black	60	.10	.63	.17
Hispanic	15	−.04	.73	−.06
Asian	34	−.09	.70	−.13
Total	610	.00	.65	

Note. Refer to Table 1 for definitions of scales. SJI = situational judgment inventory; Black = African American, Asian = Asian American.
[a] The overall means of the biodata scales were all near zero because responses were standardized before computing composites to remove the effects of elaboration. Positive values of *d* favor the non-White group, and negative values of *d* favor the White group.
* $p < .05$.

Table 10 indicates, many of these subgroup comparisons are not statistically significant because of the relatively small numbers in some (e.g., only 15 in the Hispanic group), and one should interpret all subgroup mean-difference trends with some caution.

Discussion

This article has detailed the development of biodata and SJI measures that may provide useful data in the college admissions process for either selection or student development purposes. We also described data that were collected on these measures, SAT/ACT scores, existing measures of the Big Five personality constructs, and outcome measures including first-year college GPA, self-ratings and peer ratings on a variety of student performance dimensions, and class absenteeism from 644 college freshmen. Student performance was considered broadly in terms of 12 intellectual, interpersonal, and intrapersonal dimensions, and predictors and outcomes were constructed accordingly.

Validity of Experimental Measures

Relationships between the new measures and our set of outcomes indicated potentially useful levels of criterion-related and incremental validity over and above SAT/ACT scores and existing measures of the Big Five. At the level of zero-order correlations, several of the biodata scales were correlated above .20 with GPA,

absenteeism, and peer ratings of the 12 student performance measures. The SJI composite correlated significantly only with the students' self-rating. Efforts to develop and cross-validate empirical scoring keys for both biodata and SJI were successful, producing criterion-related validities in the .40s and .50s for self-ratings and in the .20s and .30s for the other three outcomes. As expected, multivariate analyses showed that the increment in validity was higher for the outcomes (peer and self-ratings) that matched the dimensions we hoped to represent in our biodata and SJI measures than it was for GPA, but it was also of comparable magnitude for the class absence measure. Even in the case of GPA, the incremental change in the squared multiple correlation (.062) is likely to be of practical utility in most college admissions contexts.

Subgroup Differences

One concern regarding standardized tests of cognitive ability has been the sizable mean differences in the performance of minority groups, so our analyses included examining the mean scores of racial and gender subgroups. Subgroup mean differences on our new biodata and SJI measures were lower than those observed on the SAT/ACT. For the SAT/ACT racial subgroup mean differences were over one standard deviation for the African American–White and the Hispanic–White comparisons, similar to the usual reported differences. Mean differences between Asian Americans and Whites were smaller ($d = -.36$). Data on most of the biodata measures and the SJI composite indicated relatively smaller and nonsignificant racial subgroup mean differences. Differences on the GPA outcome measure indicated that all minority subgroups performed about one standard deviation below the White group. On the absenteeism and self-rating measures, subgroup means were very nearly the same with the exception that Asian Americans tended to report more absenteeism and lower self-appraisals. The number of students in all three minority groups was relatively small (Ns = 15 to 60), but the data do indicate that racial subgroup differences on the biodata and SJI composite are small or nonexistent. Mean differences on the SAT/ACT, on the other hand, are large and similar to those on the GPA outcome measure, but not on the other two outcomes for which we have data.

Female students tended to outscore male students on most of the predictors with the exception of the SAT/ACT, in which they scored lower ($d = -.29$). Women tended to outperform men on all four outcome measures (ds ranged from .11 for GPA to .39 on the self-rating measure).

Psychometric Properties of Biodata and SJI Measures

As mentioned earlier, 12 dimensions of student performance guided our construction of situational judgment and biodata measures. In the case of the biodata measures, there was reasonable evidence for the reliability and discriminability of these dimensions, though scales for at least four dimensions exhibited unacceptably low coefficient alphas. Evidence for the hypothesized dimensionality of the SJI scales was disappointing. Individual scale reliabilities were low, and there was minimal evidence for our multidimensional model of these scales. A single overall scale is the most appropriate representation of the SJI item responses, which does not tend to be an unusual finding for SJI measures in employment settings (Pulakos et al., 1994).

Biodata and SJI measures are related to existing personality measures but are almost completely uncorrelated with SAT/ACT scores. The level of relationship with personality measures, however, is not that high, and both the SJI composite and biodata added incrementally to personality measures in predicting the various outcomes considered in this study. The primary objective realized with the inclusion of these alternative potential correlates of student performance was to test whether these measures offered any predictive value beyond widely available standardized measures. Beyond this objective, the constructs underlying the biodata measures and the SJI composite are informed by their correlations with personality measures. For example, Extraversion is moderately related (i.e., above .40) with the biodata measures of Leadership and Interpersonal, Conscientiousness with Perseverance and Knowledge, and Openness with Artistic and Learning.

Limitations and Future Work

Additional psychometric development is called for in the case of both biodata and SJI measures, but perhaps the major concern in making either set of measures operational is the ability to coach students to provide answers that will inflate their scores. Both sets of measures were correlated with self-deception and impression-management measures of social desirability. Whether this correlation is performance related (i.e., students who are better at impression management also do better in their academic work) or this is a form of bias has been debated in the employment literature; most likely both are partial explanations. We think the relevant question is not whether students can be coached to get better scores on these tests (we believe they can) but whether the higher scores invalidate the utility of these measures and whether we can take steps to minimize the effects from coaching or faking these measures. We experimented with requiring some examinees in our study to elaborate in writing on some of their multiple-choice responses to biodata questions (Schmitt et al., 2003). Our major conclusion in that study was that elaboration on all items in these measures, including those that might not be so verifiable from other sources, may be successful in minimizing the impact of social desirability in biodata measures. However, the impact of elaboration may also be partially due to the fact that elaborated items were more objective and verifiable by their nature than were the items for which elaboration was not required. Much additional work in this area is required.

Another rather simple recommendation is that our work be replicated in a predictive validity study. In the case of this data collection effort, we relied on test takers who were already university students to provide the responses we analyzed. This produced some restriction of range especially on the SAT/ACT measure, which was used as one criterion to admit these students into the university. There may also be important motivational differences between our examinees and applicants for college admission. We paid our students to carry out this task, we monitored it carefully, and we believe that students were serious participants, but they were not being evaluated for college admission based on their scores. Collecting and analyzing additional data from a wider variety of students in other parts of the country and from a more diverse student group actually applying for college admission

would be a prerequisite to the operational use of these measures. Certainly, it is also necessary that large numbers of diverse student groups be examined to determine if the subgroup differences observed in this study are replicated. Related to implementing the biodata and SJI measures in different settings is the question of whether the measures are useful for admissions purposes or for student development purposes. For example, students receiving a high score on a college performance scale could be interpreted as already having high standing on that dimension of college performance in an absolute sense, or it could mean that relative to other applicants they are more likely on that dimension to reap the developmental benefits available during the college experience.

The generalizability of our SJI scoring key can also be researched further, both in terms of the empirical scoring key and in terms of the samples on which the test is used. The justification for developing a scoring key based on our sample of advanced undergraduate students lies primarily in the fact that they had successfully navigated their way through 2 years of the college experience and were on track to graduate (as opposed to other students who dropped out of college, for whatever reason). Thus, in this broad sense they are subject matter experts along both cognitive and noncognitive dimensions of our model of college performance. Although the graduation criterion is strongly tied to academic achievement, it is also related to other outcomes, because failing to deal with social or psychological problems successfully can prevent a timely graduation or continuation in college. We also felt that it would be inappropriate to develop a key based on responses by "true" experts in each performance dimension (e.g., faculty or highly accomplished students specializing in just one area). First, students must manage their time and resources in college to juggle multiple concerns across multiple performance areas. This implies that the absolute "best" response to a given situation (e.g., devoting all of one's time to a single class project) may not be the optimal one in conjunction with other concerns in shcool. Therefore, the sample on which our scoring key was based can be said to reflect an appropriate level of situational judgment that leads to overall, acceptable outcomes. Second, keys based on different expert groups might reflect lofty expectations that are unrealistic of the average, successful college student, where adequate growth and achievement is not equivalent to expertise. We investigated this to some extent by creating a new scoring key for the same SJI based on data from residence hall advisors, applying the same procedure we used in developing the initial scoring key. Results showed that 75% of the response options were scored in the same way on both keys, and composite scores for the keys correlated .97 (Friede et al., 2002). This finding, however, does not preclude the use of unique scoring keys by groups that have stringent or lenient expectations about student performance regarding situational judgment.

Finally, most college-admissions decisions are informed by high school GPA, and previous research documents the validity of high school GPA as a predictor of college GPA (e.g., Patelis & Camara, 1999; Willingham, 1985, 1998). In this study, high school GPA was not available; future research should attempt to assess the incremental validity of biodata and SJI measures in an analysis that includes this important predictor as well.

Conclusion

This study provides a broader view of student performance outcomes than is usual in assessing the predictability of student success in college. Our analysis of universities' goals for students is informative and interesting in its own right, but it also suggests that the capabilities of students and their performance outcomes be assessed more broadly than is often the case in studies of academic prediction. The results of this study are encouraging with respect to the predictive utility of alternative measures including biodata and SJI measures in admissions decision making. Specially developed empirical keys display criterion-related validity that rivals that of the SAT/ACT. The theory-based biodata and SJI scales have incremental validity over traditional standardized tests and display substantially smaller subgroup differences than do SAT/ACT measures. Criterion-related validity gains were realized for the traditional GPA outcome as well as for a broadened set of student outcomes collected from students and their peers. These alternative predictors are moderately related to some scales of the Big Five, but they possess incremental validity over these standardized personality measures, and because of their face validity they are likely to be more acceptable than are the more abstract personality measures. Analyses of the biodata scales showed evidence of the student performance dimensions we tried to build into the measures, but the SJI scales seemed to reflect a single dimension, perhaps representing knowledge of how to succeed in the broad set of life and academic situations facing new college students. As indicated in the previous section, much work needs to be done especially in analyzing, and perhaps minimizing, student response sets that may affect scores on these measures if, and when, they are used to make college admissions decisions.

References

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Personnel, 25,* 297–308.

Atkinson, R. C. (2001, April). *Standardized tests and access to American universities.* The 2001 Robert H. Atwell Distinguished Lecture presented at the 83rd annual meeting of the American Council on Education, Washington, DC.

Beatty, A., Greenwood, R., & Linn, R. L. (Eds.). (1999). *Myths and tradeoffs: The role of tests in undergraduate admissions.* Washington, DC: National Academy Press.

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

Borman, W. C., & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*(2), 99–109.

Breland, H. M. (1998). *National trends in the use of test scores in college admissions.* Unpublished manuscript.

Brown, S. H. (1981). Validity generalization and situational moderation in the life insurance industry. *Journal of Applied Psychology, 66,* 664–670.

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 258–299). 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* (pp. 35–70). San Francisco: Jossey-Bass.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey-

Schmidt, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86,* 410–417.

Clifton, T. C., Mumford, M. D., & Baughman, W. A. (1999). Background data and autobiographical memory: Effects of item types and task characteristics. *International Journal of Selection and Assessment, 7,* 57–71.

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual.* Odessa, FL: Psychological Assessment Resources.

Cress, C. M., Astin, H. S., Zimmer-Oster, K., & Burkhardt, J. (2001). Developmental outcomes of college students' involvement in leadership activities. *Journal of College Student Development, 42,* 15–27.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86,* 122–133.

Friede, A. J., Gillespie, M. A., Kim, B. H., Oswald, F. L., Ramsay, L. J., & Schmitt, N. (2002). *Final report: Development and validation of alternative measures of college success.* Princeton, NJ: College Board.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, the Netherlands: Tilburg University Press.

Goldman, R. D., & Slaughter, R. E. (1976). Why college grade point average is difficult to predict. *Journal of Educational Psychology, 68,* 9–14.

Gough, H. G., & Arvey, R. D. (1998). *Personnel Reaction Blank.* Palo Alto, CA: Consulting Psychologists Press.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

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.

Hezlett, S. A., Kuncel, N. R., Vey, M. A., Ahart, A. M., Ones, D. S., Campbell, J. P., & Camara, W. (2001, April). The predictive validity of the SAT: A meta-analysis. In D. Ones & S. Hezlett (Chairs), *Predicting performance: The interface of I-O psychology and educational research.* Symposium conducted at the 16th Annual Convention of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues, evidence, and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194.

Huffcutt, A. I., Weekley, J. A., Wiesner, W. H., DeGroot, T. G., & Jones, C. (2001). Comparison of Situational and Behavior Description Interview questions for higher-level positions. *Personnel Psychology, 54,* 619–644.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–88.

Jöreskog, K., & Sörbom, K. (2001). *LISREL 8.50*: *User's reference guide.* Chicago: Scientific Software International.

LePine, J. A., Erez, A., & Johnson, D. E. (2002). The nature and dimensionality of organizational citizenship behavior: A critical review and meta-analysis. *Journal of Applied Psychology, 87,* 52–65.

Mael, F. A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44,* 763–792.

McCrae, R. R., & Costa, P. T. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 139–153). New York: Guilford Press.

McDaniel, M. A., Bruhn-Finnegan, E. B., Morgeson, F. P., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests. *Journal of Applied Psychology, 86,* 730–740.

McGinty, S. M. (1997). *The college application essay.* New York: College Entrance Examination Board.

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., Russell, T. L., Carter, G. W., & Dunnette, M. D. (1988). *Revision of the Management Selection Interview: Final report.* Minneapolis, MN: Personnel Decisions Research Institute.

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.

Mumford, M. D. (2001). *Rationally-keyed biographical data items for constructs related to sales performance.* Fairfax, VA: George Mason University.

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3, pp. 61–138). Palo Alto, CA: Consulting Psychologists Press.

Mumford, M. D., Stokes, G. S., & Owens, W. A. (1990). *Patterns of life adaptation: The ecology of human individuality.* Hillsdale, NJ: Erlbaum.

Nickels, B. J. (1994). The nature of biodata. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook* (pp. 1–16). Palo Alto, CA: Consulting Psychologists Press.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660–679.

Owens, W. A., Albright, L. E., & Glennon, J. R. (1966). *A catalog of life history items.* Chicago: Creativity Research Institute and Richardson Foundation.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons. *Journal of Applied Psychology, 64,* 569–607.

Patelis, T., & Camara, W. (1999). *Preliminary report of statewide assessments in the United States.* New York: College Board.

Paulhus, D. L. (1988). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding.* Unpublished manual, University of British Columbia, Vancouver, British Columbia, Canada.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). New York: Academic Press.

Payne, D. A., Rapley, F. E., & Wells, R. A. (1973). Application of a biographical data inventory to estimate college academic achievement. *Measurement and Evaluation in Guidance, 6,* 152–156.

Pulakos, E. D., Arad, S., Donovan, M. A., & Plamondon, K. E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology, 84,* 612–624.

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241–258.

Pulakos, E. D., Schmitt, N., & Keenan, P. A. (1994). *Validation and implementation of the FBI Special Agent entry level selection system* (No. FR-PRO-94–20). Alexandria, VA: HumRRO.

Ra, J. B. (1989). Validity of a new evaluative scale to aid admissions decisions. *Evaluation and Program Planning, 12,* 195–204.

Regents of the University of California v. Bakke, 438 U.S. 265, 1978.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. R. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

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.

Russell, C. J., Green, S. G., & Griggs, D. (n.d.). *Biographical questionnaire items designed to discriminate between transitional and transformational leaders.* West Lafayette, IN: Purdue University.

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.

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.

Schmitt, N., Oswald, F. L., Kim, B. H., Gillespie, M. A., Ramsay, L. J., & Yoo, T. (2003). Impact of elaboration on social desirability and the validity of biodata. *Journal of Applied Psychology, 6,* 979–988.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. P. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Schmitt, N., & Kunce, C. (2002). The effect of required elaboration of answers to noncognitive measures. *Personnel Psychology, 55,* 569–588.

Stark, S., Chernyshenko, O. S., Chan, K. Y., & Lee, W. C. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86,* 943–953.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., et al. (2000). *Practical intelligence in everyday life.* New York: Cambridge University Press.

Stokes, G. S., & Cooper, L. A. (2001). Context/construct approaches in life history form development for selection. *International Journal of Selection and Assessment, 9,* 138–151.

Stokes, G. S., Mumford, M. D., & Owens, W. A. (1989). Life history prototypes in the study of human individuality. *Journal of Personality, 57,* 509–545.

Taber, T. D., & Hackman, J. D. (1976). Dimensions of undergraduate college performance. *Journal of Applied Psychology, 61,* 546–558.

University of Pennsylvania, Graduate School of Education. (2000). *Reporting on issues in education reform* (CPRE Publication No. RB-31-June). Consortium for Policy Research in Education.

Viswesvaran, C., Ones, D. S., & Hough, L. M. (2001). Do impression management scales in personality inventories predict managerial performance ratings? *International Journal of Selection and Assessment, 9,* 277–289.

Wiggins, J. S., & Trapnell, P. D. (1997). Personality structure: The return of the Big Five. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 737–765). San Diego, CA: Academic Press.

Wightman, L. F., & Jaeger, R. M. (1998). *High stakes and ubiquitous presence: An overview and comparison of Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability.* New York: College Entrance Examination Board.

Willingham, W. W. (1998, December). *Validity in college selection: Context and evidence.* Paper presented at the Workshop on the Role of Tests in Higher Education Admissions, Washington, DC.

(*Appendixes follow*)

Appendix A

Sample Biodata Items

Knowledge

Think about the last several times you have had to learn new facts or concepts about something. How much did you tend to learn?

a. usually not enough

b. sometimes not enough

c. just what is needed

d. a little more than what is needed

e. much more than what is needed

Leadership

How many times in the past year have you tried to get someone to join an activity in which you were involved or leading?

a. never

b. once

c. twice

d. three or four times

e. five times or more

Citizenship

How often have you signed a petition for something you believe in?

a. very often

b. often

c. sometimes

d. seldom

e. never

Ethics

If you were leaving a concert and noticed that someone left their purse behind with no identification, what would you do?

a. make an effort to find the person in the area, then turn the purse and its contents over to a charity if you fail

b. make an effort to find the owner; if you fail, keep the cash in the purse for yourself and give the purse to a friend

c. keep the cash and the purse

d. turn the purse over to the facility's lost and found

Appendix B

Sample Situational Judgment Inventory (SJI) Items

Knowledge

Your grade for a particular class is based on three exams, with no class attendance requirement. All of the homework requirements for the class are posted on the professor's Web site. What would you do?

a. Attend class for as long as you feel that it is helping your grades.

b. Do all the homework but only go to some of the lectures. It's the exams that count.

c. Go to all the classes anyway. The professor may say something important.

d. Skip classes, but if you did poorly on the first exam, start going to classes.

e. There is no need to go to classes. Just get the homework done, and pass the exams.

What are you most likely to do?
What are you least likely to do?

Artistic

There is a concert coming up that you think will be fantastic. No one you know is interested in going with you. What would you do?

a. Go by yourself and find someone else at the concert that went alone.

b. Try to find someone else to go with you, but if you cannot then you would not go.

c. Ask your best friend to go even if you knew that he/she wasn't as excited as you were.

d. Get two tickets and offer a free ticket to anyone you know that might want to go.

What are you most likely to do?
What are you least likely to do?

Leadership

An important class project you have been working on with a group of other students is not developing as it should because of petty differences and the need of some members to satisfy their own agenda. How would you proceed?

a. Try to solve the group problems before starting on the work.

b. Work hard by yourself to make sure the project is finished, taking on others' share of the work if necessary.

c. Talk to the professor and get suggestions about solving the problem. If that doesn't work, try to switch groups or have an independent project.

d. Schedule a number of meetings, forcing the group to interact.

e. Take charge and delegate tasks to each person. Make them responsible for their part of the project.

f. Talk to the group and demand that they start working together.

What are you most likely to do?
What are you least likely to do?

Health

In the summer and fall, you walked to class and participated in various outdoor sports. When cold weather came, you took the bus and no longer participated in sports. You find that you are gaining weight. What action would you take?

a. Participate in indoor sports and start working out indoors.

b. Try not to eat as much or eat different kinds of food.

c. Walk to classes more, go to the gym and watch what you eat.

d. Work out in your room.

e. Talk to an expert in diets and see if you can find someone who will encourage you to start working out again.

What are you most likely to do?
What are you least likely to do?

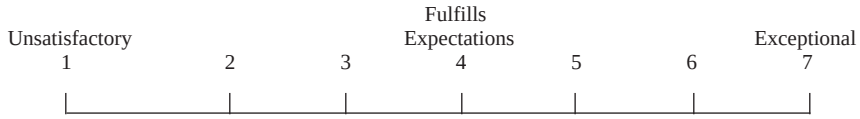
(*Appendixes continue*)

Appendix C

Sample Behaviorally Anchored Rating Scale (BARS) Items

Knowledge, Learning, Mastery of General Principles

Definition: Gaining knowledge and mastering facts, ideas, and theories and how they interrelate, and the relevant contexts in which knowledge is developed and applied. Grades or grade point average can indicate, but not guarantee, success on this dimension.



Before you make your rating, please read these two examples.

Example 1

You have never been very good at writing essays or papers and find that many of your classes in college require written assignments. You get failing grades on your first two essays even though you spent a great deal of time preparing these papers, and you realize that your classes require three more papers this term. How do you expect you would deal with this situation?

Unsatisfactory	Fulfills expectations	Exceptional
You continue with existing skill level, and hope to get better at writing by the end of the course.	You keep practicing writing essays alone, and make progress on future assignments.	You go to talk to the professors and commit to submitting extra work so that you can receive extra feedback. You make use of the writing center to learn how to write better essays.

Example 2

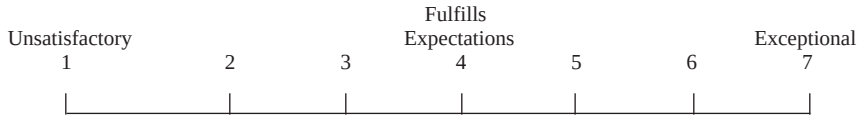
The professor has asked each member of the class to write a paper on foreign relations policy. Students are free to select different countries as the focus of their papers. What do you expect you would do?

Unsatisfactory	Fulfills expectations	Exceptional
You find someone else who is going to cover the same country, and split the work.	You choose the country about which you already have some background knowledge, and build on that.	You select a country that you know little or nothing about, and do extensive research so that you can learn from the experience.

129. Now, on your scantron, fill in the number corresponding to your level on this dimension.

Physical and Psychological Health

Definition: Possessing the physical and psychological health required to engage actively in a scholastic environment. This would include participating in healthy behaviors, such as eating properly, exercising regularly, and maintaining healthy personal and academic relations with others, as well as avoiding unhealthy behaviors, such as alcohol/drug abuse, unprotected sex, and ineffective or counterproductive coping behaviors.



Before you make your rating, please read these two examples.

Example 1

You find that you are eating more fattening and greasy food than normal and that you have not been getting sufficient exercise. You have gained 15 pounds, but find it difficult to change your eating and exercising habits. How do you expect you would deal with this situation?

Unsatisfactory	Fulfills expectations	Exceptional
You don't worry about it. You only live once, so eat what you want.	You try to establish a regular exercise routine and focus on eating healthy foods.	You get help from someone with experience in this area, such as a health professional or nutritionist and change your eating habits. You get some friends together to exercise together. There is power in numbers.

Example 2

All of the people who live near you seem to party, drink and use drugs on weekend nights. You like most of these people, but do not want to engage in some of the behavior in which they engage. You have no one else to hang out with. How do you expect you would deal with this situation?

Unsatisfactory	Fulfills expectations	Exceptional
You continue to go along with the group and their activities.	You continue to be friends and hang out with them, but do not engage in their activities.	You join a club and find other friends, and new, healthy behavior to engage in.

136. Now, on your scantron, fill in the number corresponding to your level on this dimension.

Received November 13, 2002
Revision received March 19, 2003
Accepted May 16, 2003 ■