The Non-Cognitive Questionnaire (NCQ) is a 23-item measure assessing eight noncognitive variables that are thought to predict the performance and retention of students in college. The NCQ is widely used in research and practice. This study is a meta-analytic review of the validity of scores on the NCQ across 47 independent samples for predicting academic outcomes (N = 9,321). Across all analyses, none of the scales of the NCQ are adequate predictors of GPA or persistence in college. Based on their evaluation of the NCQ, the authors recommend against its use for research or practice.

Keywords: Non-Cognitive Questionnaire; NCQ; scale validation; noncognitive variables; college admissions; college performance

The development of valid and unbiased predictors of college success for students in general, and minority students in particular, is an immediate concern to educators, researchers, and students (Nettles, 1991; Sackett, Schmitt, Ellingson, & Kabin, 2001). Some researchers argue in favor of traditional predictors like standardized tests and grade point average (GPA; Baird, 1984; Nisbet, Ruble, & Schurr, 1982; Noble, 1989, 1991), whereas other researchers argue that measurement of noncognitive variables is needed to adequately predict which students will succeed or fail (Duran, 1986; Pantages & Creedon, 1978; Sedlacek, 2004; Tracey & Sedlacek, 1984). The
effectiveness of traditional predictors of college performance, such as standardized tests, has been well illustrated (e.g., Bridgeman, McCamley-Jenkins, & Ervin, 2000; Kuncel, Credé, & Thomas, in press; Kuncel et al., 2005; Kuncel, Hezlett, & Ones, 2001, 2004; Noble, 1991), but less research has focused on noncognitive predictors.

Three goals have led to the applied interest in noncognitive predictors: increased minority admissions, improved prediction of student performance, and increased college retention of all students, but minorities in particular. All three are critical issues for higher education. Concern about the reduced admission rates of minority students that are due to the mean differences between minority and nonminority students on traditional predictors (Boyd, 1989; Wilds & Wilson, 1998) has lead to an interest in standardized noncognitive predictors to help increase minority admission rates. Similarly, researchers have also turned to noncognitive variables to understand and to address the differential rate of retention for minority and nonminority students. In particular, retention rates of African American students are a great deal lower than the retention rates of European American and Asian American students (e.g., Astin, 1999; Bynum & Thompson, 1999). Recently, large-scale efforts have been directed at developing and evaluating new noncognitive predictors of performance, retention, and degree attainment of students (e.g., Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004). For example, Le, Casillas, Robbins, and Langley (2005) developed the Student Readiness Inventory, designed to measure psychosocial and academically related skills (e.g., study skills, problem-solving skills) predictive of college success as measured by GPA and persistence.

Such inventories are based on findings that a number of noncognitive individual difference variables can be important predictors of success and can add incremental validity to traditional predictors. Indeed, the relationship between noncognitive variables (e.g., maturity, motivation, self-concept, interpersonal skills, personality variables) and non-cognitively oriented measures (e.g., biographical information, personal interviews, letters of recommendation) on college student performance has been extensively studied (e.g., Izaak, 2002; McCarthy & Goffin, 2001; Scott et al., 1995) with numerous researchers arguing that noncognitive factors are important determinants of success and persistence in college (e.g., Sedlacek, 2004; Wood, Smith, Altmair, Tarico, & Franken, 1990).

For example, Tinto’s model of student persistence outlines the manner in which individual characteristics, social and academic support, and prior qualifications impact the ability of the student to become academically and socially integrated into the university community (see Tinto, 1975, 1993). Similarly, the Adult Persistence in Learning Model (MacKinnon-Slaney, 1994) acknowledges that the manner in which college persistence is influenced, not only by academic skills and abilities (e.g., intelligence, study skills), but also by personal and socioenvironmental issues such as values, goals, interpersonal competence, and the fit between the student and the college environment. This influence of background, attitude, and environmental variables is also a key feature of Bean’s (1980) model of attrition.
Measurement of Noncognitive Variables

To some extent, colleges and universities already make use of noncognitive variables in the admissions process. Sources of noncognitive information that colleges consider when evaluating student applications include community service, letters of recommendation, personal statements, work experience, volunteer projects, and extracurricular activities. Although the use of noncognitive selection methods varies across colleges, the most common admission criteria are cognitive variables and prior grades. These variables are often given more weight in the admissions process than noncognitive measures (Moll, 1979; Willingham & Breland, 1982).

The Non-Cognitive Questionnaire (NCQ; Tracey & Sedlacek, 1984) was designed to close a perceived racial gap that exists in the admissions process. Tracey and Sedlacek (1984) argued that the difference in scores on standardized tests among African American and European American students is due to “culturally/racially biased” predictors and criteria (p. 171) and reported some evidence for relationships between the NCQ and 1st-year grades.

An evaluation of the NCQ is necessary because the NCQ has been widely used, studied, and promoted for admissions and student development purposes. The authors of the NCQ recommend the measure as an alternative admissions device, as a tool for advising students, and, in research settings, as a method of assessing the relationship between certain noncognitive variables and college student success (e.g., Sedlacek, 2004; Tracey & Sedlacek, 1984, 1987a, 1987b). For example, the NCQ is used as a predictor for admissions at Louisiana State University Medical School, North Carolina State University, and Muhlenberg College (Sedlacek, 2004). Additionally, the University of Maryland employs the NCQ in a counseling center for student development, teaching, and advising (Sedlacek, 2004; University of Maryland, 2004). Finally, the NCQ also represents one of the major bases for awarding the Gates Millennium Scholarships (Bill & Melinda Gates Foundation, 2004; Sedlacek, 2004). In all cases, decision makers wish to obtain information about students, which they hope will improve their ability to predict success.

The emphasis placed on psychosocial and motivational influences by the models of student persistence and student success reviewed earlier is reflected in the subscales of the NCQ. In addition to the theoretical support offered by the work of Tinto (1993), MacKinnon-Slaney (1994), and Bean (1985), the potential importance of each construct is also supported by varying amounts of empirical research that establish its relevance to understanding student persistence and student success. In the following sections, we discuss some of the evidence that suggests that the constructs the NCQ is designed to measure may be important predictors of college performance and persistence. All of the subscales of the NCQ are defined in Table 1.
Positive Self-Concept

Positive self-concept has repeatedly been related to success in college (e.g., Aspinwall & Taylor, 1992; Coopersmith, 1968; Epps, 1969; Eysenck & Eysenck, 1969). The association of a positive self-concept with high levels of determination, confidence, and independence (Sedlacek, 2004) increases the likelihood that students persist in their studies even in the face of difficult and/or unmet expectations.

Realistic Self-Appraisal

Realistic self-appraisal refers to a student’s recognition of his or her inadequacies and the ability to assess one’s own strengths and weaknesses (Sedlacek, 2004). A student’s realistic appraisal of his or her academic capabilities will increase the likelihood that the student engages in adequate studying and help-seeking behaviors. The accuracy of self-evaluations of ability and academic performance has been found to be positively related with subsequent performance (e.g., Eshel & Kurman, 1991; Moreland, Miller, & Laucka, 1981).

Table 1
Non-Cognitive Questionnaire (NCQ) Subscales: Eight Dimensions and Definitions

<table>
<thead>
<tr>
<th>NCQ Subscale Dimension</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Self-Concept</td>
<td>“Strong self-feeling, strength of character, determination, independence” (Sedlacek, 1987, p. 485). The score on this subscale ranges from a score of 7 to 27. The higher the score, the more desirable the characteristic.</td>
</tr>
<tr>
<td>Realistic Self-Appraisal</td>
<td>“Acceptance of personal deficiencies and tendency to engage in self-development, recognizes the need to broaden his or her uniqueness, is adapt at assessing their own performance” (Sedlacek, 1987, p. 485). The score on this subscale ranges from a score of 4 to 14.</td>
</tr>
<tr>
<td>Understanding of and Ability to Deal With Racism</td>
<td>“The acknowledgement of existing racist systems, and attempts to fight racist ideology” (Sedlacek, 1987, p. 485). The score on this subscale ranges from a score of 5 to 25.</td>
</tr>
<tr>
<td>Preference for Long-Term Goals</td>
<td>“Deferment of immediate gratification and existence of long-term educational goals” (Sedlacek, 1987, p. 485) the score on this subscale ranges from a score of 3 to 15.</td>
</tr>
<tr>
<td>Availability of Strong Support Person</td>
<td>Is an individual who has the availability of help in a crisis (Sedlacek, 1987, p. 485). The score on this subscale ranges from a score of 3 to 15.</td>
</tr>
<tr>
<td>Successful Leadership Experience</td>
<td>“Leadership experience in any area pertinent to his/her background (gang leader, church, sports, non-educational groups, etc.)” (Sedlacek, 1987, p. 485). The score on this subscale ranges from a score of 3 to 15.</td>
</tr>
<tr>
<td>Demonstrated Community Service</td>
<td>“Involvement in his/her cultural community” (Sedlacek, 1987, p. 485). The score on this subscale ranges from a score of 2 to 8.</td>
</tr>
<tr>
<td>Knowledge Acquired in a Field</td>
<td>“The individual may have unusual, nontraditional or culturally related ways of obtaining information” (Sedlacek, 1987, p. 485). The score on this subscale ranges from a score of 2 to 8.</td>
</tr>
</tbody>
</table>
Understanding and Ability to Deal With Racism

Sedlacek (1994) described this construct as reflecting a student’s ability to work within the institution’s established system in a positive manner. From the perspective of Tinto’s model (1975), this construct is particularly important in determining the degree to which minority students are able to effectively achieve social and academic integration. This skill should also be associated with satisfaction with college, and hence college retention (Bean, 1985).

Preference for Long-Term Goals

Long-term goal orientations, and particularly long-term academic goals, are likely to be adaptive in the college environment in terms of its established relationship to delay of gratification (e.g., Mischel & Ayduk, 2004). Given that goal setting for even short-term goals has a large and positive effect on subsequent performance (Locke, 1968), it would seem that setting appropriate goals is critical for students’ success. Therefore, we would expect that effective students should have distal long-term goals that result in more productive short-term goals.

Availability of Strong Support Person

College freshmen and new graduate students, especially minorities on predominately European American campuses, are not only faced with separation from friends and family but also face the challenge of adapting to a new environment and of becoming socially and academically integrated. This integration process is especially important for new college students (Bennett & Okinaka, 1989; Martin, 1990; Smith, 1991; Thomas, 1986), because unsuccessful adaptation and social adjustment has been linked to student attrition (Bragg, 1994; Kaase, 1994).

Leadership

Leadership experience is demonstrated by taking initiative and becoming involved in extracurricular activities. Lindgren (1973) examined the relationship between leadership and GPA and found strong correlations between leadership ratings and GPA (.41 for men and .38 for women). In addition, many conceptualizations of student success also include effective performance as a leader of a sports team, fraternity/sorority, club, or professional student organization (e.g., Oswald et al., 2004; Richard, Holland, & Lutz, 1967).

Community Involvement

Community involvement could be considered an important predictor of student success. Existing community involvement may be indicative of the student’s ability
to reconcile the time demands of academic and nonacademic activities and provide social networks for the student, thereby increasing the likelihood that academic and social integration are achieved in the college environment. Also, community involvement indicates likely involvement in college activities, which is often considered desirable (Astin, 1999; Conner, 1991; Lemley, 1991).

**Knowledge Acquired in a Field**

Sedlacek (2004) described this construct as the ability of the individual to learn and to acquire knowledge in less conventional ways. Sedlacek suggested that individuals, particularly minorities, may have specific, cultural knowledge obtained outside of the classroom. Astin (1975) found that minorities who exhibited knowledge that was acquired unconventionally, through credit by examination, were more likely to persist in college.

**Study Objectives**

In view of the controversy surrounding admission procedures as a whole, it is important to examine instruments such as the NCQ to ensure that universities are selecting students on the basis of valid information. Given the widespread use of the NCQ in admissions decisions and research settings, further examination of this instrument is warranted.

Our objectives are threefold. The first is establishing the predictive validity of scores on the NCQ. Second, average score differences between different racial and ethnic groups warrant study, as they would directly impact the proportion of different racial and ethnic groups that would be admitted to college. Given the objective of giving minority students an opportunity to demonstrate strengths in different domains, it would be particularly troubling if minorities scored, on average, lower on the NCQ. If minorities scored higher on the NCQ, then concerns about the fairness of the measure for admissions might also arise. Finally, predictive validity differences between the groups would be important for both applied and theoretical reasons. If the validity of scores on the NCQ is moderated by race, it would provide evidence that the determinants of performance differ in importance across racial and ethnic groups. This would be a finding of clear theoretical importance. Yet a validity difference across groups is also an indicator of predictor bias because it suggests that scores are not comparably associated with performance for both groups across the range of possible test scores (see Cascio, 1998, for a review). Overall, the results of this study could have important implications for the use of the NCQ in admissions decisions.

The purpose of this study is to use meta-analytic methods to examine the predictive validity of scores on the NCQ. In addition, this study will examine the potential moderating effect of race and gender on the validity of scores on the NCQ. Finally,
this study will examine average score differences across different racial/ethnic and gender groups.

**Method**

The NCQ was designed to measure eight noncognitive characteristics that may be valuable in admissions decisions (Tracey & Sedlacek, 1984). The NCQ contains 23 items, as follows: 18 Likert-format items; 2 multiple-choice items on educational aspirations; and 3 open-ended items pertaining to present goals, past accomplishments, and other activities. Although the authors of the NCQ often cite an estimate of the test-retest reliabilities over a 2-week interval as evidence of the measure’s temporal stability (correlations ranged from .70 to .94 across items), this estimate is based on a sample size of $N = 18$ (Tracey & Sedlacek, 1984), leaving the test-retest reliability effectively unknown. Interrater reliability on the 3 open-ended NCQ items ranged from .73 to 1.00 (Ting, 2000). From a sample of 1,766 students, internal consistency reliabilities (alpha) were reported with a range of .39 to .82 with an average of .57 (Tracey & Sedlacek, 1987b). The eight dimensions of the NCQ are presented in Table 1.

Potential data sources were identified via literature searches in the PsychINFO (1887-2002), ERIC (1966-2002), Education Full Text, and Dissertations Abstracts International (1861-2002) databases. This search was conducted to obtain all published and unpublished research (e.g., dissertations, ERIC documents). In addition, in each article or dissertation collected, the introduction, discussion, and reference section were examined for references that would contribute additional data to the study. Literature searches were conducted using the following search terms: NCQ, non-cognitive variables, non-cognitive questionnaire, non-cognitive predictors, non-intellectual variables, non-intellective predictors, and alternative predictors. Hand searches were also conducted in Educational and Psychological Measurement and the Journal of College Student Development to identify studies missed by electronic database searches.

In the end, 311 articles, dissertations, and reports were identified as potentially having relevant information. Of the 311 articles that were originally identified, 233 articles and dissertations did not have useable NCQ data. Specifically, of the 233 studies, 160 did not include any information about the NCQ, and 73 articles or dissertations were review articles not containing empirical data or using a criterion not related to college student success (e.g., counseling center use). Seventy-eight articles or dissertations did contain NCQ data; however, only 42 of these contained data that were usable. Usable data includes effect sizes (e.g., correlations, Cohen’s $d$) or statistical results from which effect sizes can be obtained (e.g., frequency tables, $\chi^2$, $t$ values). This study analyzed the standardized difference between two means (i.e., Cohen’s $d$ statistic) or the zero-order correlation between the independent and dependent variable. The remaining 36 studies were excluded because effect sizes were unobtainable from the studies or appeared to be based on data examined in
other studies. For example, if a study used multiple regression but only reported a multiple correlation without providing a predictor intercorrelation matrix, we were unable to calculate the bivariate effect sizes. Other examples of this exclusion are studies that used stepwise regression (Hood, 1992), multiple regression (Anderson, 1998), ridge regression (Tracey, Sedlacek, & Miars, 1983), and canonical correlations (White & Sedlacek, 1986). Although it is true that effect sizes can be obtained if certain information is provided in these types of analyses, too little information was present to estimate the bivariate effects. Also excluded from the analysis was a dissertation (Chung, 1996) reporting a negative alpha reliability, suggesting miskeyed test items. Given that the method of analysis used in a study is probably not correlated with the magnitude of the results, the absence of these excluded data is not likely to have a systematic bias on the results of the meta-analysis. In each of these situations, the author was contacted to obtain the original data. Unfortunately, none of the authors contacted complied with our requests.

A few authors reported only those results with statistically significant effect sizes. There are a number of different approaches to handling this problem. One approach is to include the articles that report only statistically significant results with the rest of the database. If this route were taken, the inclusion of such data would result in an upwardly biased estimate of the validity estimate. Another approach is to use statistical methods (i.e., Hedges & Olkin, 1985) to estimate the omitted effect sizes. These statistical methods require that certain assumptions be made and that the alpha level be used to filter study results. In this study, only 2% of the total database was excluded due to nonreporting of statistically nonsignificant results. Of the studies excluded for this reason, only one included the alpha value (e.g., Boyer & Sedlacek, 1988). The third approach is to include all of the missing, statistically nonsignificant effect sizes by replacing them with zeroes. This approach effectively assumes there are no Type II errors and would, on average, bias the correlations downward. The fourth approach is to exclude these studies in their entirety. Given the alternatives and small number of articles in which only statistically significant correlations were reported, the decision was made to exclude any study that failed to report findings due to significance tests.

All of the remaining articles and dissertations were coded for study relevant information by one of the three authors to extract the necessary criterion and predictor information. The reliability of coded meta-analyses is high in studies where the coding decisions are clear-cut (as is the case with this meta-analysis; Whetzel & McDaniel, 1988; Zakzanis, 1998). All of the studies were coded by one author and then coded again by a second author to ensure accuracy and reliability of the coding. In addition, the database was double-checked by the second coder to identify data entry errors. Overall, coding errors were rare and minor, with no major discrepancies between coders identified. In addition to coding for type of criterion and predictor, additional information was gathered on effect sizes, sample sizes, age, race, reported
reliabilities, validation strategies, and year in college. This additional information was collected to identify possible moderator effects. One study (Tracey & Sedlacek, 1987a) presented item-level validity coefficients and item intercorrelations. These items were organized according to scoring procedures for the NCQ published in Sedlacek (2004), and scale validities were estimated using unit-weighted composites (Nunnally, 1978). To ensure independence of samples, studies from the same authors were examined for overlap. If an overlap of samples was identified, the study with the larger sample size or more complete data was used.

The final database for the criterion-related validity meta-analyses included 47 independent samples and a total of 490 correlations, with a total pairwise sample size of 9,321. The final database for the meta-analysis of standardized mean differences \((d)\) included 14 independent samples and a total of 54 effect sizes with a total pairwise sample size of 3,298.

Nearly all of the data were for college outcomes. The exceptions are one sample of 1st-year dental school students for which correlations with 1st-year dental school performance were used, one sample of 1st-year medical students for which correlations with prior undergraduate GPA were used, and one sample of physical therapy students for which correlations with overall student GPA were used. These three studies contributed to the analyses for the GPA criterion. Given the proximity in time between early education in graduate school and the similarity between taking college courses and early graduate courses, we felt that these situations closely approximated performance in college and fit with the overall objective of evaluating the validity of scores on the NCQ for academic performance. As it turns out, inclusion or exclusion of these data from the analyses has little effect on the results. Most studies used had predictive designs; however, three studies contained data where NCQ scores collected just after college were correlated with prior grades.

The Hunter and Schmidt (1990, 2004) meta-analytic procedure was used to combine all relevant study results and to correct for artifacts that bias the relationship between the predictor and the criterion. This meta-analytic method has been used in education to examine the validity of several standardized tests used in higher education admissions including the Millers Analogies Test, Graduate Record Examination, Graduate Management Admissions Test, and Pharmacy College Admissions Test (Kuncel et al., 2001, 2005, in press; Kuncel et al., 2004).

When there is disagreement in the literature that is based primarily on statistically significant or statistically nonsignificant results, meta-analyses can provide a quantitative summary of multiple studies to reach a conclusion. The Hunter and Schmidt method (1990, 2004) is a random-effects method that provides a way to estimate the parameters of the distribution of effect sizes (i.e., correlations) across different studies. In other words, the Hunter and Schmidt method permits estimation of the extent to which the variability observed in the literature (i.e., differently sized correlations across studies) is due to sampling error or other artifactual sources of study variance.
First, the total variance of the observed correlations is estimated, and the variance due to artifacts is subtracted to give the residual variance. Because many studies in this area employ smaller sample sizes, the ability to examine this variability can aid us in deciding if differences across studies are due to substantive moderators (e.g., the type of college) or are merely artifacts of small sample studies. A meta-analysis calculates the sample size weighted mean correlation obtained from the individual study effect sizes. The 90% credibility intervals and the $SD_{\text{obs}}$ (i.e., the observed standard deviation before removing variance due to statistical artifacts) and $SD_\rho$ (the standard deviation of true validity after removing variance due to statistical artifacts) are all associated with the variability in correlations across studies. The $SD_\rho$ is an estimate of effect size heterogeneity after accounting for statistical artifacts. The $SD_\rho$ is used to calculate the 90% credibility interval, which provides information about the range of possible values the correlations might take across situations. A lower 90% credibility interval greater than zero indicates a likely positive relationship between the predictor (i.e., NCQ subscales) and the criterion (i.e., GPA, persistence, credits earned) across situations, although the actual correlations may vary. This is also termed validity generalization. In this study, most of the effects are nearly zero and credibility intervals also include zero, indicating that the NCQ subscale scores are not valid across situations.

### Results

The individual validities of the subscales of the NCQ with regard to GPA, persistence, and credits earned are found in Tables 2, 3, and 4, respectively. The validities for the subscales of the NCQ and GPA for the overall sample are low across all the

<table>
<thead>
<tr>
<th>Subscale</th>
<th>$N$</th>
<th>$k$</th>
<th>$r_{\text{obs}}$</th>
<th>$SD_{\text{obs}}$</th>
<th>$SD_\rho$</th>
<th>90% Cred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-concept</td>
<td>7,703</td>
<td>27</td>
<td>-.05</td>
<td>.12</td>
<td>.10</td>
<td>-.21 to .11</td>
</tr>
<tr>
<td>Self-appraisal</td>
<td>7,681</td>
<td>29</td>
<td>.06</td>
<td>.08</td>
<td>.05</td>
<td>-.02 to .14</td>
</tr>
<tr>
<td>Racism</td>
<td>7,175</td>
<td>28</td>
<td>.01</td>
<td>.09</td>
<td>.07</td>
<td>-.11 to .13</td>
</tr>
<tr>
<td>Long-term goals</td>
<td>7,523</td>
<td>28</td>
<td>.06</td>
<td>.09</td>
<td>.06</td>
<td>-.04 to .16</td>
</tr>
<tr>
<td>Support person</td>
<td>8,124</td>
<td>30</td>
<td>.04</td>
<td>.07</td>
<td>.03</td>
<td>-.01 to .09</td>
</tr>
<tr>
<td>Leadership</td>
<td>7,016</td>
<td>27</td>
<td>.04</td>
<td>.08</td>
<td>.06</td>
<td>-.06 to .14</td>
</tr>
<tr>
<td>Community service</td>
<td>7,172</td>
<td>28</td>
<td>.07</td>
<td>.08</td>
<td>.04</td>
<td>.00 to .14</td>
</tr>
<tr>
<td>Nontraditional knowledge</td>
<td>5,430</td>
<td>22</td>
<td>.07</td>
<td>.07</td>
<td>.03</td>
<td>.02 to .12</td>
</tr>
</tbody>
</table>

Note: $N$ = number of subject; $k$ = number of studies; $r_{\text{obs}}$ = sample size weighted mean observed correlation; $SD_{\text{obs}}$ = observed standard deviation; $SD_\rho$ = standard deviation of true validity; 90% cred. = 90% credibility interval.
subscales, and range from $r_{obs} = -0.05$ for the Self-Concept subscale to $r_{obs} = 0.07$ for the Community Service and Nontraditional Knowledge subscales. The validities found for the persistence criterion for the overall sample range from $r_{obs} = 0.14$ for the Self-Concept subscale to $r_{obs} = -0.08$ for the Nontraditional Knowledge subscale (Table 3). Similarly, the validities found for the criterion of credits earned range from $r_{obs} = 0.04$ for the Support Person subscale to $r_{obs} = 0.15$ for the Long-Term Goals subscale (Table 4).

The validities for European and African Americans with regards to the subscales of the NCQ and their relationship with GPA are found in Tables 5 and 6, respectively. The results found for the subscales of the NCQ and GPA, for European Americans,
again show low validities (Table 5). The validities found range from $r_{\text{obs}} = -0.11$ for the Self-Concept subscale to $r_{\text{obs}} = 0.07$ for the Self-Appraisal and Long-Term Goals subscales. Similarly, the validities found for African Americans are found in Table 6. The validities for the subscales range from $r_{\text{obs}} = 0.04$ for the Support Person subscale to $r_{\text{obs}} = 0.13$ for the Long-Term Goals subscale.

The individual validities for the NCQ subscales for the overall female and male samples are found in Table 7 and 8. The validities found for the female sample range from $r_{\text{obs}} = -0.05$ for the Self-Concept subscale to $r_{\text{obs}} = 0.10$ for the Long-Term Goals and Support Person subscales. For males, the validities are found in Table 8. Again, the NCQ shows low validities ranging from a $r_{\text{obs}} = -0.07$ for Self-Concept subscale to a $r_{\text{obs}} = 0.10$ for the Community Service subscale.

**Table 5**

Meta-Analyses for GPA Criterion for European American Students

<table>
<thead>
<tr>
<th>Subscale</th>
<th>N</th>
<th>k</th>
<th>$r_{\text{obs}}$</th>
<th>$SD_{\text{obs}}$</th>
<th>$SD_\rho$</th>
<th>90% Cred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-concept</td>
<td>5,378</td>
<td>9</td>
<td>-0.11</td>
<td>0.08</td>
<td>0.06</td>
<td>-0.21 to -0.01</td>
</tr>
<tr>
<td>Self-appraisal</td>
<td>5,361</td>
<td>8</td>
<td>0.07</td>
<td>0.06</td>
<td>0.04</td>
<td>0.00 to 0.14</td>
</tr>
<tr>
<td>Understanding racism</td>
<td>5,035</td>
<td>7</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01 to 0.07</td>
</tr>
<tr>
<td>Long-term goals</td>
<td>5,361</td>
<td>8</td>
<td>0.07</td>
<td>0.09</td>
<td>0.08</td>
<td>-0.06 to 0.20</td>
</tr>
<tr>
<td>Support person</td>
<td>5,361</td>
<td>8</td>
<td>0.04</td>
<td>0.03</td>
<td>0.00</td>
<td>0.04 to 0.04</td>
</tr>
<tr>
<td>Leadership</td>
<td>5,035</td>
<td>7</td>
<td>0.04</td>
<td>0.08</td>
<td>0.07</td>
<td>-0.08 to 0.16</td>
</tr>
<tr>
<td>Community service</td>
<td>5,035</td>
<td>7</td>
<td>0.06</td>
<td>0.04</td>
<td>0.02</td>
<td>0.03 to 0.09</td>
</tr>
<tr>
<td>Nontraditional knowledge</td>
<td>3,489</td>
<td>5</td>
<td>0.06</td>
<td>0.05</td>
<td>0.03</td>
<td>0.01 to 0.10</td>
</tr>
</tbody>
</table>

Note: $N$ = number of subject; $k$ = number of studies; $r_{\text{obs}}$ = sample size weighted mean observed correlation; $SD_{\text{obs}}$ = observed standard deviation; $SD_\rho$ = standard deviation of true validity; 90% cred. = 90% credibility interval.

**Table 6**

Meta-Analyses for GPA Criterion for African American Students

<table>
<thead>
<tr>
<th>Subscale</th>
<th>N</th>
<th>k</th>
<th>$r_{\text{obs}}$</th>
<th>$SD_{\text{obs}}$</th>
<th>$SD_\rho$</th>
<th>90% Cred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-concept</td>
<td>1,171</td>
<td>10</td>
<td>0.02</td>
<td>0.12</td>
<td>0.08</td>
<td>-0.11 to 0.15</td>
</tr>
<tr>
<td>Self-appraisal</td>
<td>1,115</td>
<td>9</td>
<td>0.04</td>
<td>0.05</td>
<td>0.00</td>
<td>0.04 to 0.04</td>
</tr>
<tr>
<td>Racism</td>
<td>1,115</td>
<td>9</td>
<td>0.04</td>
<td>0.13</td>
<td>0.10</td>
<td>-0.12 to 0.20</td>
</tr>
<tr>
<td>Long-term goals</td>
<td>956</td>
<td>8</td>
<td>0.13</td>
<td>0.10</td>
<td>0.04</td>
<td>0.06 to 0.20</td>
</tr>
<tr>
<td>Support person</td>
<td>1,016</td>
<td>9</td>
<td>0.01</td>
<td>0.10</td>
<td>0.03</td>
<td>-0.04 to 0.06</td>
</tr>
<tr>
<td>Leadership</td>
<td>956</td>
<td>8</td>
<td>0.03</td>
<td>0.07</td>
<td>0.00</td>
<td>0.03 to 0.03</td>
</tr>
<tr>
<td>Community service</td>
<td>956</td>
<td>8</td>
<td>0.09</td>
<td>0.06</td>
<td>0.00</td>
<td>0.09 to 0.09</td>
</tr>
<tr>
<td>Nontraditional knowledge</td>
<td>873</td>
<td>7</td>
<td>0.08</td>
<td>0.10</td>
<td>0.05</td>
<td>-0.00 to 0.16</td>
</tr>
</tbody>
</table>

Note: $N$ = number of subject; $k$ = number of studies; $r_{\text{obs}}$ = sample size weighted mean observed correlation; $SD_{\text{obs}}$ = observed standard deviation; $SD_\rho$ = standard deviation of true validity; 90% cred. = 90% credibility interval.
Inter correlations between the scales were reported for six studies and were combined meta-analytically and are presented in Table 9. The correlations between the scales of the NCQ range from .03 to .42 based on a sample size of 641. The largest relationship was between Leadership and Community Service (r = .42).

Mean differences on NCQ scores between African American and European American students were reported in a number of studies. Meta-analytic estimates of the standardized mean difference are presented in Table 10. Negative effect sizes indicate that, on average, African American students obtained higher scores. The mean differences between racial groups on the NCQ range from \(d = -0.52\) for the Racism subscale to \(d = 0.01\) for the Support Person subscale.

### Table 7
**Meta-Analyses for GPA Criterion for Female Students**

<table>
<thead>
<tr>
<th>Scale</th>
<th>N</th>
<th>k</th>
<th>(r_{obs})</th>
<th>SD_{obs}</th>
<th>SD_{p}</th>
<th>90% Cred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-concept</td>
<td>1,372</td>
<td>4</td>
<td>-.05</td>
<td>.09</td>
<td>.07</td>
<td>-.16 to .07</td>
</tr>
<tr>
<td>Self-appraisal</td>
<td>1,246</td>
<td>3</td>
<td>.08</td>
<td>.03</td>
<td>.00</td>
<td>.08 to .08</td>
</tr>
<tr>
<td>Racism</td>
<td>1,246</td>
<td>3</td>
<td>.04</td>
<td>.06</td>
<td>.02</td>
<td>.01 to .07</td>
</tr>
<tr>
<td>Long-term goals</td>
<td>1,246</td>
<td>3</td>
<td>.10</td>
<td>.04</td>
<td>.00</td>
<td>.10 to .10</td>
</tr>
<tr>
<td>Support person</td>
<td>1,459</td>
<td>4</td>
<td>.10</td>
<td>.01</td>
<td>.00</td>
<td>.10 to .10</td>
</tr>
<tr>
<td>Leadership</td>
<td>1,246</td>
<td>3</td>
<td>.08</td>
<td>.01</td>
<td>.00</td>
<td>.08 to .08</td>
</tr>
<tr>
<td>Community service</td>
<td>1,246</td>
<td>3</td>
<td>.06</td>
<td>.03</td>
<td>.00</td>
<td>.06 to .06</td>
</tr>
<tr>
<td>Nontraditional knowledge</td>
<td>1,246</td>
<td>3</td>
<td>.04</td>
<td>.04</td>
<td>.00</td>
<td>.04 to .04</td>
</tr>
</tbody>
</table>

Note: \(N = \) number of subject; \(k = \) number of studies; \(r_{obs} = \) sample size weighted mean observed correlation; \(SD_{obs} = \) observed standard deviation; \(SD_{p} = \) standard deviation of true validity; 90% cred. = 90% credibility interval.

### Table 8
**Meta-Analyses for GPA Criterion for Male Students**

<table>
<thead>
<tr>
<th>Scale</th>
<th>N</th>
<th>k</th>
<th>(r_{obs})</th>
<th>SD_{obs}</th>
<th>SD_{p}</th>
<th>90% Cred.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-concept</td>
<td>1,521</td>
<td>3</td>
<td>-.07</td>
<td>.04</td>
<td>.00</td>
<td>-.07 to -.07</td>
</tr>
<tr>
<td>Self-appraisal</td>
<td>1,521</td>
<td>3</td>
<td>.07</td>
<td>.04</td>
<td>.00</td>
<td>.07 to .07</td>
</tr>
<tr>
<td>Racism</td>
<td>1,521</td>
<td>3</td>
<td>.01</td>
<td>.03</td>
<td>.00</td>
<td>.01 to .01</td>
</tr>
<tr>
<td>Long-term goals</td>
<td>1,521</td>
<td>3</td>
<td>.08</td>
<td>.05</td>
<td>.00</td>
<td>.08 to .08</td>
</tr>
<tr>
<td>Support person</td>
<td>1,750</td>
<td>4</td>
<td>.05</td>
<td>.05</td>
<td>.00</td>
<td>.05 to .05</td>
</tr>
<tr>
<td>Leadership</td>
<td>1,521</td>
<td>3</td>
<td>.06</td>
<td>.04</td>
<td>.00</td>
<td>.06 to .06</td>
</tr>
<tr>
<td>Community service</td>
<td>1,521</td>
<td>3</td>
<td>.10</td>
<td>.02</td>
<td>.00</td>
<td>.10 to .10</td>
</tr>
<tr>
<td>Nontraditional knowledge</td>
<td>1,521</td>
<td>3</td>
<td>.07</td>
<td>.01</td>
<td>.00</td>
<td>.07 to .07</td>
</tr>
</tbody>
</table>

Note: \(N = \) number of subject; \(k = \) number of studies; \(r_{obs} = \) sample size weighted mean observed correlation; \(SD_{obs} = \) observed standard deviation; \(SD_{p} = \) standard deviation of true validity; 90% cred. = 90% credibility interval.
Discussion

Our results suggest that NCQ scores are largely unrelated to college performance as measured by GPA, college persistence, and credits earned. The NCQ as a whole was not found to be a valid selection tool for admissions purposes. Assuming colleges and universities are concerned about predicting grades, college persistence, and credits earned, our results suggest that the NCQ should not be used for admissions decisions. For the persistence criterion, only one correlation greater than .10 was obtained. The GPA analyses included a sufficient number of studies to place confidence in the 90% credibility intervals, and for all scales, the intervals are narrow and include zero, suggesting that the validity of the NCQ scores probably do not vary much from zero from situation to situation.

Table 9
Meta-Analysis of the Intercorrelations Among Non-Cognitive Questionnaire (NCQ) Scales

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Positive self-concept</td>
<td>.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Realistic self-appraisal</td>
<td>.16</td>
<td>.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Racism</td>
<td>.28</td>
<td>.12</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Long-term goals</td>
<td>.06</td>
<td>.15</td>
<td>.17</td>
<td>.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Availability of support</td>
<td>.17</td>
<td>.10</td>
<td>.20</td>
<td>.18</td>
<td>.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Leadership</td>
<td>.04</td>
<td>.10</td>
<td>.12</td>
<td>.11</td>
<td>.23</td>
<td>.42</td>
<td>.19</td>
</tr>
<tr>
<td>7. Community service</td>
<td>.07</td>
<td>.03</td>
<td>.05</td>
<td>.22</td>
<td>.12</td>
<td>.37</td>
<td></td>
</tr>
<tr>
<td>8. Acquired knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N = 641, k = 6 for all correlations.

Table 10
Non-Cognitive Questionnaire (NCQ) Mean Differences Between African American and European American Students

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>k</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-concept</td>
<td>3,298</td>
<td>7</td>
<td>−.27</td>
</tr>
<tr>
<td>Self-appraisal</td>
<td>3,298</td>
<td>7</td>
<td>−.07</td>
</tr>
<tr>
<td>Racism</td>
<td>3,298</td>
<td>7</td>
<td>−.52</td>
</tr>
<tr>
<td>Long-term goals</td>
<td>3,298</td>
<td>7</td>
<td>−.33</td>
</tr>
<tr>
<td>Support person</td>
<td>3,298</td>
<td>7</td>
<td>.01</td>
</tr>
<tr>
<td>Leadership</td>
<td>3,298</td>
<td>7</td>
<td>−.18</td>
</tr>
<tr>
<td>Community service</td>
<td>3,298</td>
<td>7</td>
<td>.00</td>
</tr>
<tr>
<td>Nontraditional knowledge</td>
<td>3,176</td>
<td>5</td>
<td>−.12</td>
</tr>
</tbody>
</table>

Note: N = sample size; k = number of studies; d = standardized mean difference. Positive ds favor European American students, negative ds favor African American students.
Interestingly, the mean differences suggest that, on average, African American students score somewhat higher on the NCQ than do European American students. As a result, use of the NCQ will tend to increase admission of minority students. However, students admitted on the basis of NCQ scores are only trivially more likely to succeed than if they had been selected at random. Given these characteristics, this approach approximates near-random assignment of bonus points for race in an admissions decision. That is, in admissions decisions using the NCQ, African American students will, on average, receive a better rank order position when being evaluated by the NCQ, yet, due to the poor validity of the scores on the NCQ, this improved position appears to have little to do with individual difference information other than race. Given recent court decisions (e.g., *Gratz v. Bollinger*, 2003; *Grutter v. Bollinger*, 2003), in our opinion, this seems to be a questionable and risky approach.

It is important to note that the initial validation study (Tracey & Sedlacek, 1984) and several studies that followed have incorrectly concluded that the NCQ is a valid predictor of GPA and persistence in college (Ancis & Sedlacek, 1997; Fuertes, Sedlacek, & Liu, 1994). In one study (Ancis & Sedlacek, 1997), stepwise multiple regression was used to examine cumulative GPA over a seven-semester period. However, only some of the NCQ subscales predicted GPA across each semester, and none of the subscales predicted GPA for all semesters, suggesting that the associations are likely chance relationships. In another study (Fuertes et al., 1994), stepwise multiple discriminant analysis was used to examine persistence over seven semesters. Again, only some of the NCQ subscales predicted college persistence across each semester, and none of the subscales predicted college persistence for all semesters.

These studies illustrate a general issue. It is our belief that some scholars in this area have been mislead by stepwise multiple regression, stepwise discriminant analysis, and other related methods. In fact, any set of random variables will produce a positive multiple $R$ (Weisberg, 1985) due to capitalization of small chance effects unless their correlations with the outcome are precisely zero. When they are filtered by significance levels (as in a stepwise regression), large and spurious relationships can be observed as seen in Cureton’s (1950) classic and entertaining illustration. The fact that a few NCQ scales generally demonstrate very small positive correlations (.10–.15) would only increase the magnitude of the multiple $R$, leading to the conclusion that the NCQ is a useful predictor. However, this appearance of validity is anchored in optimally estimated weights that capitalize on small chance variations in the data. This is particularly the case when specific scales are allowed to be included in a model on the basis of a significance test and negative weights rather than a priori hypotheses based on theory. Even random numbers will explain some of the variance because you never get exactly zero for a correlation in a real research setting. Because of this, stepwise methods are particularly fraught with peril. At each step, the stepwise procedure adds the most statistically significant predictor term (i.e., the one with the lowest $p$ value) until there are none left that meet criteria to enter the equations.
The few studies that suggest more favorable findings (e.g., Tracey & Sedlacek, 1985) report that half or more of the scales fail to make a significant incremental contribution to predictive validity despite large sample sizes and despite the fact that the sample consisted of entering freshmen (i.e., demonstrating representativeness). In other words, this study found that when the subscales were added to the SAT, only four of the subscales added to the prediction of 1st-year GPA for African Americans. Again, this approach to research risks capitalizing on chance. With enough scales, some will be included in the model and demonstrate effects just by chance. Yet these variables are added without regard to theory and are weighted either positively or negatively depending on what optimizes the model even if the original scale was expected to have the opposite relation with the criterion. These weights can be highly unstable and do not result in an estimate of test validity that would be typically obtained using the scales (Freedman, 1983; Rencher & Pun, 1980). We calculated the validity of the scores on the entire NCQ as a unit-weighted composite using the meta-analytic data presented here (Nunnally, 1978). This estimates the validity of the scores on the entire measure when combined without differential weighting and using each scale score in the proposed direction. The validity of the scores on the entire measure for predicting GPA when combined in this manner is a mere .09.

Psychometric Problems

It is important to note that we do not believe the lack of validity of the scores on the NCQ is necessarily due to unimportant constructs. Two large reviews of the literature found that personality, interest, and other noncognitive variables were useful predictors of student performance (Breland, 1981; Lavin, 1965). It is our suspicion that the lack of validity of the scores on the NCQ is due to the measurement properties of the NCQ. In other words, if the constructs were operationalized differently, better results might be obtained. As it stands, the NCQ has only two or three items for some of the scales with generally poor internal consistency.

An instrument that has low reliability cannot be expected to provide valid information about the behavior it is intended to measure. Internal consistency is a measure of item homogeneity. When items on a test are combined into a composite score the consistency of the items reflects the ability to interpret the composite scores. Effect sizes are attenuated by the reliability of scores on a test. In other words, if the scores on a test are unreliable, then the observed correlations with important outcomes like GPA will be lower (Reinhardt, 1996; Vacha-Haase, Henson, & Caruso, 2002). Specifically regarding the scores on the NCQ, reliability is important when the interpretation of scores led to college admission outcomes. The more measurement error in the scores, the less useful the scores become. Items on each of the subscales of the NCQ should correlate highly with each other if they are truly assessing the same construct. The low reliability of the scores on the NCQ restricts the degree
of observed validity that is possible. Nunnally and Bernstein (1994) suggested that internal consistency estimates for college admissions decisions should reach at least a reliability of .90 or greater. The internal consistency reliabilities of the scores on the NCQ averaged .57 (Tracey & Sedlacek, 1987b). Furthermore, ratings for the open-ended items could be better constructed by using behavioral anchors, which would also improve their measurement properties.

**Item Structure and Scoring Problems**

Some items are not true dominance items. Instead, strong endorsement could mean either positive or negative things about a student. For example, strongly disagreeing to the item “I am as skilled academically as the average applicant to this school” or “My high school grades don’t really reflect what I can do” could mean that the student perceives himself or herself as either especially competent or especially incompetent. Finally, the NCQ scoring of open-ended items indicates that points should be received for nonresponse (Sedlacek, 2004). For example, listing no goals whatsoever receives a higher score than listing an immediate goal. With all that is known about goal setting, this is a poor scoring key. Nonresponse, particularly in such a short measure, should lead to invalidating the scale for that test taker rather than a positive score.

**Item Content Problems**

Finally, we also believe that some of the NCQ items in the measure are not good measures of the construct of interest. For example, Realistic Self-Appraisal items that ask about a student’s perception of his or her ability relative to others or the ease with which the student can obtain a grade at a particular institution do not, in isolation, reflect realistic self-appraisal. Assessing a student’s recognition and acceptance of deficiencies requires information about a student’s actual standing relative to peers or the actual difficulty of the institution relative to the student’s current skill level. In their current form, the meaning of item responses is ambiguous.

**Conclusion**

Research suggests that increasing the number of minority students in college can have positive outcomes for people of all races (Bowen & Bok, 1998). One approach to increasing minority enrollment and improving the quality of admission decisions is the development of additional valid predictors that measure important characteristics not captured by current admissions measures. Indeed, recent measure development research appears to have advanced the area of noncognitive predictors (e.g., Le et al., 2005; Oswald et al., 2004). However, the NCQ does not appear to be one of them. The consistently poor performance of the NCQ indicates that the NCQ is
not a valid predictor and should not be used to make decisions that affect students’ lives. Overall, we admire the ends sought by the developers and users of the NCQ but do not admire the means used to pursue them.

References

References marked with an asterisk indicate studies included in the meta-analysis.


*Fuller, M. K. (1994). The contribution of selected cognitive and noncognitive variables to the academic success of medical technology students. Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.


