Cognitively loaded tests of knowledge, skill, and ability often contribute to decisions regarding education, jobs, licensure, or certification. Users of such tests often face difficult choices when trying to optimize both the performance and ethnic diversity of chosen individuals. The authors describe the nature of this quandary, review research on different strategies to address it, and recommend using selection materials that assess the full range of relevant attributes using a format that minimizes verbal content as much as is consistent with the outcome one is trying to achieve. They also recommend the use of test preparation, face-valid assessments, and the consideration of relevant job or life experiences. Regardless of the strategy adopted, it is unreasonable to expect that one can maximize both the performance and ethnic diversity of selected individuals.

Cognitively loaded tests of knowledge, skill, and ability are commonly used to help make employment, academic admission, licensure, and certification decisions (D’Costa, 1993; Dwyer & Ramsey, 1995; Frierson, 1986; Mehrens, 1989). Law school applicants submit scores on the Law School Admission Test (LSAT) for consideration when making admission decisions. Upon graduation, the same individuals must pass a state-administered bar exam to receive licensure to practice. Organizations commonly rely on cognitive ability tests when making entry-level selection decisions and tests of knowledge and skill when conducting advanced-level selection. High-school seniors take the Scholastic Assessment Test (SAT) for use when determining college admissions and the distribution of scholarship funds. Testing in these settings is termed high stakes, given the central role played by such tests in determining who will and who will not gain access to employment, education, and licensure or certification (jointly referred to as credentialing) opportunities.

The use of standardized tests in the knowledge, skill, ability, and achievement domains for the purpose of facilitating high-stakes decision making has a history characterized by three dominant features. First, extensive research has demonstrated that well-developed tests in these domains are valid for their intended purpose. They are useful, albeit imperfect, descriptors of the current level of knowledge, skill, ability, or achievement. Thus, they are meaningful contributors to credentialing decisions and useful predictors of future performance in employment and academic settings (Mehrens, 1999; Neisser et al., 1996; Schmidt & Hunter, 1998; Wightman, 1997; Wilson, 1981). Second, racial group differences are repeatedly observed in scores on standardized knowledge, skill, ability, and achievement tests. In education, employment, and credentialing contexts, test score distributions consistently reveal significant mean differences by race (e.g., Bobko, Roth, & Potosky, 1999; Hartigan & Wigdor, 1989; Jensen, 1980; Lynn, 1996; Neisser et al., 1996; Scarr, 1981; Schmidt, 1988; N. Schmitt, Clause, & Pulakos, 1996; Wightman, 1997; Wilson, 1981). Blacks tend to score approximately one standard deviation lower than Whites, and Hispanics score approximately two thirds of a standard deviation lower than Whites. Asians typically score higher than Whites on measures of mathematical-quantitative ability and lower than Whites on measures of verbal ability and comprehension. These mean differences in test scores can translate into large adverse impact against protected groups when test scores are used in selection and credentialing decision making. As subgroup mean differences in test scores increase, it becomes more likely that a smaller proportion of the lower scoring subgroup will be selected or granted a credential (Sackett & Wilk, 1994).

Third, the presence of subgroup differences leads to questions regarding whether the differences observed bias
resulting decisions. An extensive body of research in both the employment and education literatures has demonstrated that these tests generally do not exhibit predictive bias. In other words, standardized tests do not underpredict the performance of minority group members (e.g., American Educational Research Association, American Psychological Association, National Council of Measurement in Education, 1999; Cole, 1981; Jensen, 1980; Neisser et al., 1996; O'Connor, 1989; Sackett & Wilk, 1994; Wightman, 1997; Wilson, 1981).

These features of traditional tests cause considerable tension for many organizations and institutions of higher learning. Most value that which is gained through the use of tests valid for their intended purpose (e.g., a higher performing workforce, a higher achieving student body, a cadre of credentialed teachers who meet knowledge, skill, and achievement standards). Yet, most also value racial and ethnic diversity in the workforce or student body, with rationales ranging from a desire to mirror the composition of the community to a belief that academic experiences or workplace effectiveness are enhanced by exposure to diverse perspectives. What quickly becomes clear is that these two values—performance and diversity—come into conflict. Increasing emphasis on the use of tests in the interest of gaining enhanced performance has predictable negative consequences for the selection of Blacks and Hispanics. Conversely, decreasing emphasis on the use of tests in the interest of achieving a diverse group of selectees often results in a substantial reduction in the performance gains that can be recognized through test use (e.g., Schmidt, Mack, & Hunter, 1984; N. Schmitt et al., 1996).

This dilemma is well-known, and a variety of resolution strategies have been proposed. One class of strategies involves some form of minority group preference; these strategies were the subject of an American Psychologist article by Sackett and Wilk (1994) that detailed the history, rationale, consequences, and legal status of such strategies. However, a variety of recent developments indicate a growing trend toward bans on preference-based forms of affirmative action. The passage of the Civil Rights Act of 1991 made it unlawful for employers to adjust test scores as a function of applicants’ membership in a protected group. The U.S. Supreme Court’s decisions in City of Richmond v. J. A. Croson Co. (1989) and Adarand Constructors, Inc. v. Pen (1995) to overturn set-aside programs that reserved a percentage of contract work for minority-owned businesses signaled the Court’s stance toward preference-based affirmative action (Mishkin, 1996). The U.S. Fifth Circuit Court of Appeals ruled in Hopwood v. State of Texas (1996) that race could not be used as a factor in university admissions decisions (Kier & Davenport, 1997; Mishkin, 1996). In 1996, the state of California passed Proposition 209 prohibiting the use of group membership as a basis for any selection decisions made by the state, thus affecting public sector employment and California college admissions (Pear, 1996). Similarly, state of Washington voters approved Initiative 200 that bars the use of race in state hiring, contracting, and college admissions (Verhovek & Ayres, 1998).

**Strategies for Achieving Diversity Without Minority Preference**

In light of this legal trend toward restrictions on preference-based routes to diversity, a key question emerges: What are the prospects for achieving diversity without minority preference and without sacrificing the predictive accuracy and content relevancy present in knowledge, skill, ability, and achievement tests? Implicit in this question is the premise that one values both diversity and the performance outcomes that an organization or educational institution may realize through the use of tests. If one is willing to sacrifice quality of measurement and predictive accuracy, there are many routes to achieving diversity including random selection, setting a low cut score, or the use of a low-impact predictor even though it may possess little to no predictive power. On the other hand, if one values performance outcomes but does not value diversity, maximizing predictive accuracy can be the sole focus. We suggest that most organizations and educational institutions espouse neither of these extreme views and instead seek a balance between diversity concerns and performance outcomes. Clearly, the use of traditional tests without race-based score adjustment fails to achieve such a balance. However, what alternatives are available for use in high-stakes, large-scale assessment contexts? In this article, we review various alternative strategies that have been put forth in the employment, education, and credentialing literatures. We note that some strategies have been examined more carefully in some domains than in others, and thus the attention we devote to employment, education, and credentialing varies across these alternatives.

The first strategy involves the measurement of constructs with little or no adverse impact along with traditional cognitively loaded knowledge, skill, ability, and
achievement measures. The notion is that if we consider other relevant constructs along with knowledge, skill, ability, and achievement measures when making high-stakes decisions, subgroup differences should be lessened because alternatives such as measures of interpersonal skills or personality usually exhibit smaller differences between ethnic and racial subgroups. A second strategy investigates test items in an effort to identify and remove those items that are culturally laden. It is generally believed that because those items likely reflect irrelevant, culture-bound factors, their removal will improve minority passing rates. The use of computer or video technology to present test stimuli and collect examinee responses constitutes a third strategy. Using these technologies usually serves to minimize the reading and writing requirements of a test. Reduction of adverse impact may be possible when the reading or writing requirements are inappropriately high. Also, video technology may permit the presentation of stimulus materials in a fashion that more closely matches the performance situation of interest. Attempts to modify how examinees approach the test-taking experience constitutes a fourth strategy. To the extent that individuals of varying ethnic and racial groups exhibit different levels of test-taking motivation, attempts to enhance examinee motivation levels may reduce subgroup differences. Furthermore, changing the way in which the test and its questions are presented may impact how examinees respond, a result that could also facilitate minority test performance. A fifth strategy has been to document relevant knowledge, accomplishments, or achievements via portfolios, performance assessments, or accomplishment records. Proponents of this strategy maintain that this approach is directly relevant to desired outcomes and hence should constitute a more fair assessment of the knowledge, skill, ability, or achievement domain of interest for members of all subgroups. Finally, we also review the use of coaching or orientation programs that provide examinees with information about the test and study materials or aids to facilitate optimal performance. In addition, we consider whether modifying the time limits prescribed for testing helps reduce subgroup differences. In the following sections, we review the literature relevant to each of these efforts in order to understand the nature of subgroup differences on knowledge, skill, ability, and achievement tests and to ascertain the degree to which these efforts have been effective in reducing these differences.

Use of Measures of Additional Relevant Constructs

Cognitively loaded knowledge, skill, ability, and achievement tests are among the most valid predictors available when selecting individuals across a wide variety of educational and employment situations (Schmidt & Hunter, 1981, 1998). Therefore, a strategy for resolving the dilemma that allows for the use of such tests is readily appealing. To that end, previous research has identified a number of noncognitive predictors that are also valid when making selection decisions in most educational and employment contexts. Measures of personality and interpersonal skills generally exhibit smaller mean differences by ethnicity and race and also are related to performance on the job or in school (e.g., Barrick & Mount, 1991; Bobko et al., 1999; Mount & Barrick, 1995; Sackett & Wilk, 1994; Salgado, 1997; N. Schmitt et al., 1996; Wolfe & Johnson, 1995). The use of valid, noncognitive predictors, in combination with cognitive predictors, serves as a very desirable strategy in that it offers the possibility of simultaneously meeting multiple objectives. If additional constructs, beyond those measured by the traditional test, are relevant for the job or educational outcomes of interest, supplementing cognitive tests offers the prospect of increased validity when predicting those outcomes. If those additional constructs are ones on which subgroup differences are smaller, a composite of the traditional test and the additional measures will often exhibit smaller subgroup differences than the traditional test alone. The prospect of simultaneously increasing validity and reducing subgroup differences makes this a strategy worthy of careful study.

Several different approaches have been followed when examining this strategy. On the basis of the psychometric theory of composites, Sackett and Ellingson (1997) developed a set of implications helpful in estimating the effect of a supplemental strategy on adverse impact. First, consider a composite of two uncorrelated measures, where \( d_1 = 1.0 \) and \( d_2 = 0.0 \). Although intuition may suggest that a composite of the two will split the difference (i.e., result in a \( d \) of 0.5), the computed value is 0.71. Thus, whereas supplementing a cognitively loaded test with an uncorrelated measure exhibiting no subgroup differences will reduce the composite subgroup difference, this reduction will be less than some might expect. Second, a composite may result in a \( d \) larger than either of the components making up the composite if the two measures are moderately corre-
Each additional measure is uncorrelated with the original outcome one hopes to predict; see N. Schmitt, Rogers, and 11.45, respectively. (Note that adding additional predictors reduced subgroup differences by 0.24 standard deviation units and increased the multiple correlation by 0.13. These results mirror earlier research conducted by Ones, Viswesvaran, and Schmidt (1993), in which integrity test validities were estimated and discussed with respect to their influence on adverse impact via composite formation.

Weighting of criterion components. Each of these studies is predicated on the assumption that the outcome measure against which the predictors are being validated is appropriately constituted. Educational performance, job performance, and credentialing criteria, however, may be defined in terms of a single dimension or multiple dimensions. For example, if the outcome of interest involves only the design of a new product or performance on an academic test, performance may be unidimensional. If, however, the outcome of interest requires citizenship behavior or effective coordination as well, then performance will be more broadly defined. When performance is multidimensional, institutions may choose to assign different weights to those various dimensions in an effort to reflect that they vary in importance.

De Corte (1999) and Hattrup, Rock, and Scalia (1997) have shown how the weighting of different elements of the criterion space can affect the regression weights assigned to different predictors, the level of adverse impact, and predicted performance. Hattrup et al. used cumulated correlations from three different studies to estimate the relationship between contextual performance (behaviors supporting an organization’s climate and culture), task performance (behaviors that support the delivery of goods or services), cognitive ability, and work orientation. Using regression analyses, cognitive ability and work orientation were used to predict criterion composites in which contextual performance and task performance were assigned varying weights. The regression weight for cognitive ability was highest when task performance was weighted heavily in the criterion composite, whereas work orientation received the largest regression weight when contextual performance was the more important part of the criterion. As expected, adverse impact was the greatest when task performance was weighted more heavily and it decreased as contextual performance received more weight. Relative to a composite wherein task and contextual performance were weighted equally, the percentage of minorities that would be selected varied considerably depending on the relative weights.
The traditional focus of studies examining differential item functioning (DIF; Berk, 1982) has been on the identification of items that function differently for minority versus majority test takers. Conceivably, such items would contribute to misleading test scores for members of a particular subgroup. Statistically, DIF seeks items that vary in difficulty for members of subgroups that are actually evenly matched on the measured construct. That is, an attempt is made to identify characteristics of items that lead to poorer performance for minority-group test takers than for equally able majority-group test takers. Assuming such a subset of items or item characteristics exist, they must define a race-related construct that is distinguishable from the construct the test is intended to measure (McCauley &endoza, 1985). Perhaps because of the availability of larger sample sizes in large-scale testing programs, much of the DIF research conducted to date has been done using educational and credentialing tests.

Initial evidence for DIF was provided by Medley and Quirk (1974), who found relatively large group by item interactions in a study of the performance of Black and White examinees on the National Teacher Examination items reflecting African American art, music, and literature. One should note, however, that these results are based on examinee performance on a test constructed using culturally specific content. Ironson and Subkoviak (1979) also found evidence for DIF when they evaluated five cognitive subtests administered as part of the National Longitudinal Study of 1972. The verbal subtest, measuring vocabulary and reading comprehension, contained the largest number of items flagged as biased. Items at the end of each of the subtests also tended to be biased against Black examinees, presumably because of lower levels of reading ability or speed in completing these tests. More recently, T. Freedle and Kostin (1990; R. Freedle & Kostin, 1997) showed that Black examinees were more likely to get difficult verbal

Summary. Research on the strategy of measuring different relevant constructs illustrates that it does matter which individual differences are assessed in high-stakes decision making if one is concerned about maximizing minority subgroup opportunity. Minority pass rates can be improved by including noncognitive predictors in a test battery. However, adding predictors with little or no impact will not eliminate adverse impact from a battery of tests that includes cognitively loaded knowledge, skill, ability, and achievement measures. Reduction in adverse impact results from a complex interaction between the validity of the individual predictors, their intercorrelation, the size of subgroup differences on the combination of tests used, the selection ratio, and the manner in which the tests are used. In fact, in most situations wherein a variety of knowledge, skills, and abilities are considered when making selection decisions, adverse impact will remain at legally unacceptable levels and subgroup mean differences on the predictor battery will not be a great deal lower than the differences observed for cognitive ability alone. The composition of the test battery should reflect the individual differences required to perform in the domain of interest. If institutions focus mainly or solely on task performance, then cognitive ability will likely be the most important predictor and adverse impact will be great. If, however, they focus on a broader domain that involves motivational, interpersonal, or personality dimensions as well as cognitive ability, then adverse impact may be reduced.

Identification and Removal of Culturally Biased Test Items

A second strategy pursued in an attempt to resolve the performance versus diversity dilemma involves investigating the possibility that certain types of test items are biased. The Hattrup et al. (1997) and De Corte (1999) analyses were formulated in employment terms, yet the same principles hold for educational and credentialing tests as well. For example, when licensing lawyers, the licensing body is concerned about both technical competence and professional ethics. The general principle relevant across settings is that when multiple criterion dimensions are of interest, the weights given to the criterion dimensions can have important effects on the relationship between the predictors and the overall criterion. The higher the weight given to cognitively loaded criterion dimensions, the higher the resulting subgroup difference. In response, one may be tempted to simply choose criterion dimension weights on the basis of their potential for reducing subgroup differences. Such a strategy would be errant, however, as criterion weights should be determined primarily on the basis of an analysis of the performance domain of interest and the values that an institution places on the various criterion dimensions.

Summary. Research on the strategy of measuring different relevant constructs illustrates that it does matter which individual differences are assessed in high-stakes decision making if one is concerned about maximizing minority subgroup opportunity. Minority pass rates can be improved by including noncognitive predictors in a test battery. However, adding predictors with little or no impact will not eliminate adverse impact from a battery of tests that includes cognitively loaded knowledge, skill, ability, and achievement measures. Reduction in adverse impact results from a complex interaction between the validity of
items on the Graduate Record Examination (GRE) and the SAT correct when compared with equally able White examinees. However, the Black examinees were less likely to get the easy items right. In explanation, they suggested that the easier items possessed multiple meanings more familiar to White examinees, whose culture was most dominant in the test items and the educational system.

Whitney and Schmitt (1997) investigated the extent to which DIF may be present in biographical data measures developed for use in an employment context. In an effort to further DIF research, Whitney and Schmitt focused not only on identifying whether biodata items may exhibit DIF, but also whether the presence of DIF can be traced to differences in cultural values between racial subgroups. More than one fourth of the biodata items exhibited DIF between Black and White examinees. Moreover, the Black and White examinees differentially endorsed item response options designed to reflect differing cultural notions of human nature, the environment, time orientation, and interpersonal relations. However, there was only limited evidence that the differences observed between subgroups in cultural values were actually associated with DIF. After removal of all DIF items, the observed disparity in test scores between Blacks and Whites was eliminated, a disparity that, incidentally, favored Blacks over Whites.

Other studies evaluating the presence of DIF have proved less interpretable. Scheuneman (1987) developed item pairs that reflected seven major hypotheses about potential sources of item bias in the experimental portion of the GRE. The results were inconsistent in showing that these manipulations produced a more difficult test for minority versus majority examinees. In some instances, the manipulations produced a greater impact on Whites than Blacks. In other instances, a three-way interaction between group, test version, and items indicated that some uncontrolled factor (e.g., content of a passage or item) was responsible for the subgroup difference. Scheuneman and Gerritz (1990) examined verbal items from the GRE and the SAT that consisted of short passages followed by questions. Although they did identify several item features possibly linked to subgroup differences (e.g., content dealing with science, requiring that examinees identify the major thesis in a paragraph), the results, as a whole, yielded no clear-cut explanations. Scheuneman and Gerritz concluded that DIF may result from a combination of item features, the most important of which seems to be the content of the items. Similarly inconclusive results were reported by A. P. Schmitt and Dorans (1990) in a series of studies on SAT-Verbal performance. Items that involved the use of homographs (i.e., words that are spelled like other words with a different meaning) were more difficult for otherwise equally able racial and ethnic group members. Yet, when nonnative English speakers were removed from the analyses, there were few remaining DIF items.

Schmeiser and Ferguson (1978) examined the English usage and social studies reading tests of the American College Test (ACT) and found little support for DIF. Two English tests and three social studies tests were developed to contain different content, while targeting the same cognitive skills. None of the interactions between test content and racial and ethnic group were statistically significant. Similarly, Scheuneman and Grima (1997) reported that the verbal characteristics of word problems (e.g., readability indexes, the nature of the arguments, and propositions) in the quantitative section of the GRE were not related to DIF indexes.

These results indicate that although DIF may be detected for a variety of test items, it is often the case that the magnitude of the DIF effect is very small. Furthermore, there does not appear to be a consistent pattern of items favoring one group versus another. Results do not indicate that removing these items would have a large impact on overall test scores. In addition, we know little about how DIF item removal will affect test validity. However, certain themes across these studies suggest the potential for some DIF considerations. Familiarity with the content of items appears to be important. The verbal complexity of the items is also implicated, yet it is not clear what constitutes verbal complexity. Differences in culture are often cited as important determinants of DIF; but beyond the influence of having English as one’s primary language, we know little about how cultural differences play a role in test item performance.

**Use of Alternate Modes of Presenting Test Stimuli**

A third strategy to reduce subgroup differences in tests of knowledge, skill, ability, and achievement has been to change the mode in which test items or stimulus materials are presented. Most often this involves using video or auditory presentation of test items, as opposed to presenting test items in the normal paper-and-pencil mode. Implicit in this strategy is the assumption that reducing irrelevant written or verbal requirements will reduce subgroup differences. In support of this premise, Sacco et al. (2000) demonstrated the relationship between reading level and subgroup differences by assessing the degree to which the readability level of situational judgment tests (SJTs) was correlated with the size of subgroup differences in SJT performance. They estimated that 10th-, 12th-, and 14th-grade reading levels would be associated with Black–White ds of 0.51, 0.62, and 0.74, respectively; reading levels at the 8th, 10th, and 13th grades would be associated with Hispanic–White ds of 0.38, 0.48, and 0.58, respectively. This would suggest that reducing the readability level of a test should in turn reduce subgroup differences. One must be careful, however, to remove only verbal requirements irrelevant to the criterion of interest, as Sacco et al. also demonstrated that verbal ability may partially account for SJT validities. In reviewing the research on this strategy, we focused on three key studies investigating video as an alternative format to a traditional test. The three selected studies illustrate issues central in examining the effects of format changes on subgroup differences. For other useful studies that also touch on this issue, see Weekly and Jones (1997) and N. Schmitt and Mills (in press).

Pulakos and Schmitt (1996) examined three measures of verbal ability. A paper-and-pencil measure testing verbal...
analyses, vocabulary, and reading comprehension had a validity of .19 when predicting job performance and a Black–White $d = 1.03$. A measure that required examinees to evaluate written material and write a persuasive essay on that material had a validity of .22 and a $d = 0.91$. The third measure of verbal ability required examinees to draft a description of what transpired in a short video. The validity of this measure was .19 with $d = 0.45$. All three measures had comparable reliabilities (i.e., .85 to .92). Comparing the three measures, there was some reduction in $d$ when written materials involving a realistic reproduction of tasks required on the job were used rather than a multiple-choice paper-and-pencil test. When the stimulus material was visual rather than written, there was a much greater reduction in subgroup differences (i.e., Black–White $d$ values dropped from 1.03 to 0.45, with parallel findings reported for Hispanics). In terms of validity, the traditional verbal ability test was the most predictive ($r = .39$), whereas the video test was less so ($r = .29$, corrected for range restriction and criterion unreliability).

Chan and Schmitt (1997) evaluated subgroup differences in performance on a video-based SJT and on a written SJT identical to the script used to produce the video enactment. The written version of the test displayed a subgroup difference of 0.95 favoring Whites over Blacks, whereas the video-based version of the test produced a subgroup difference of only 0.21. Corrections for unreliability produced differences of 1.19 and 0.28. These differences in $d$ were matched by subgroup differences in perceptions of the two tests. Both groups were more favorably disposed (as indicated by perceptions of the tests’ face validity) to the video version of the test, but Blacks significantly more so than Whites.

Sackett (1998) summarized research on the use of the Multistate Bar Examination (MBE), a multiple-choice test of legal knowledge and reasoning, and research conducted by Klein (1983) examining a video-based alternative to the MBE. The video test presented vignettes of lawyers taking action in various settings. After each vignette, examinees were asked to evaluate the actions taken. Millman, Mehrens, and Sackett (1993) reported a Black–White $d$ of 0.89 for the MBE; an identical value ($d = 0.89$) was estimated by Sackett for the video test.

Taken at face value, the Sackett (1998), Chan and Schmitt (1997), and Pulakos and Schmitt (1996) results appear contradictory regarding the impact of changing from a paper-and-pencil format to a video format. These contradictions are more apparent than real. First, it is important to consider the nature of the focal construct. In the Pulakos and Schmitt study of verbal ability and the Sackett study of legal knowledge and reasoning, the focal construct was clearly cognitive in nature, falling squarely into the domain of traditional knowledge, skill, ability and achievement tests. In Chan and Schmitt, however, the focal construct (i.e., interpersonal skill) was, as the name implies, not heavily cognitively loaded. In fact, the SJT used was of the type often suggested as a potential additional measure that might supplement a traditional test. Chan and Schmitt provided data supporting the hypothesis that the relatively large drop in $d$ was due in part to the removal of an implicit reading comprehension component present in the written version of the SJT. When that component was removed through video presentation, the measure became less cognitively loaded and $d$ was reduced.

But what of the differences between the verbal skill (Pulakos & Schmitt, 1996) and legal knowledge and reasoning studies (Sackett, 1998)? We use this comparison to highlight the importance of examining the correlation between the traditional written test and the alternative video test. In the legal knowledge and reasoning study, in which the video did not result in reduced subgroup differences, the correlation between the traditional test and the video test was .89, corrected for unreliability. This suggests that the change from paper-and-pencil to video was essentially a format change only, with the two tests measuring the same constructs. In the verbal skills study, in which the video resulted in markedly smaller subgroup differences, the correlation between the two, corrected for unreliability in the two measures, was .31. This suggests that the video-based test in the verbal skills study reflected not only a change in format, but a change in the constructs measured as well. An examination of the scoring procedures for the verbal skills video test supports this conclusion. Examinee essays describing video content were rated on features of verbal ability (e.g., sentence structure, spelling) and on completeness of details reported. We suggest that scoring for completeness introduced into the construct domain personality characteristics such as conscientiousness and detail orientation, both traits exhibiting smaller subgroup differences. Consistent with the arguments made above with regard to the ability of composites to reduce subgroup differences, the reduction in $d$ observed for the verbal skills video test was due not to the change in format, but rather to the introduction of additional constructs that ameliorated the influence of verbal ability when determining $d$. Thus, the research to date indicates that changing to a video format does not per se lead to a reduction in subgroup differences. Future research into the influence of alternative modes of testing should take steps to control for the unintended introduction of additional constructs beyond those being evaluated. Failure to separate test content from test mode will confound results, blurring our ability to understand the actual mechanism responsible for reducing subgroup differences.

Last, we wish to highlight the role of reliability when comparing traditional tests with alternatives. Focusing on the legal knowledge and reasoning study (Sackett, 1998), recall that the Black–White difference was identical ($d = 0.89$) for both tests. We now add reliability data to our discussion. Internal consistency reliability for the MBE was .91; correcting the subgroup difference for unreliability results in a corrected value of $d = 0.93$. Internal consistency reliability for the video test was .64; correcting the subgroup difference for unreliability results in a corrected value of $d = 1.11$. In other words, after taking differences in reliability into account, the alternative video test results in a larger subgroup difference than the traditional paper-and-pencil test. Such an increase is possible in...
that a traditional test may tap declarative knowledge, whereas a video alternative may require the application of that knowledge, an activity that will likely draw on higher order cognitive skills resulting in higher levels of d. Clearly, any conclusion about the effects of changing test format on subgroup differences must take into account reliability of measurement. What appears to be a format effect may simply be a reliability effect. Because different reliability estimation methods focus on different sources of measurement error (e.g., inconsistency across scorers, content sampling), taking reliability into account will require considering the most likely sources of measurement error in a particular setting.

The results reported in this section document the large differences in d that are typically observed when comparisons are made between different test formats. Examinations of reasons for these differences are less conclusive. It is not clear that the lower values of d are a function of test format. Cognitive ability requirements and the reading requirements of these tests seem to be a major explanation. Clearly more research would be helpful. Separating the method of assessment from the content measured is a major challenge when designing studies to evaluate the impact of test format. However, it is a challenge that must be surmounted in order to understand whether group differences can be reduced by altering test format. Future research should also consider investigating how a change in format influences validity. Only one of the three studies reviewed here evaluated validity effects; clearly, that issue warrants additional attention.

**Use of Motivation and Instructional Sets**

A fourth strategy places the focus not on the test itself, but on the mental state adopted by examinees and its role in determining test performance. An individual’s motivation to complete a test has the potential to influence test performance. To the extent that there are racial group differences in test-taking motivation, energizing an individual to persevere when completing a test may reduce subgroup differences. At a more general level, a test taker’s attitude and approach toward testing may also influence test performance. Such test-taking impressions could be partially determined by the instructional sets provided when completing tests and the context in which actual test questions are derived. Manipulating instructional sets and item contexts has the potential to alter observed group differences if these aspects of the testing experience allow individuals of varying racial groups to draw on culture-specific cognitive processes.

A number of studies have demonstrated racial group differences in test-taking motivation. O'Neil and Brown (1997) found that eighth-grade Black students reported exerting the least amount of effort when completing a math exam. Hispanic students reported exerting slightly more, whereas White students reported exerting the most effort when completing the exam. Chan, Schmitt, DeShon, Clause, and Delbridge (1997) demonstrated that the relationship between race and test performance was partially mediated by test-taking motivation, although the mediating effect accounted for a very small portion of the variance in test performance. In a follow-up study, Chan, Schmitt, Sacco, and DeShon (1998) found that pretest reactions affected test performance and mediated the relationship between belief in tests and test performance. Their subgroup samples were not large, but motivational effects operated similarly for Whites and Blacks.

As a test of the influence of item context, a unique study by DeShon, Smith, Chan, and Schmitt (1998) investigated whether presenting test questions in a certain way would reduce subgroup differences. They tested the hypothesis proposed by Helms (1992) that cognitive ability tests fail to adequately assess Black intelligence because they do not account for the emphasis in Black culture on social relations and social context, an observation offered at a more general level by others as well (e.g., Miller-Jones, 1989; O'Connor, 1989). Contrary to Helms’s argument, racial subgroup performance differences on a set of Wason conditional reasoning problems were not reduced by presenting the problems in a social relationship form.

Another working hypothesis is that the mere knowledge of cultural stereotypes may affect test performance. In other words, making salient to test takers their ethnic and racial or their gender identity may alter both women’s and minorities’ test-taking motivation, self concept, effort level, and expectation of successful performance. Steele and colleagues (Steele, 1997; Steele & Aronson, 1995) proposed a provocative theory of stereotype threat that suggests that the way in which a test is presented to examinees can affect examinee performance. The theory hypothesizes that when a person enters a situation wherein a stereotype of the group to which that person belongs becomes salient, concerns about being judged according to that stereotype arise and inhibit performance. When members of racial minority groups encounter high-stakes tests, their awareness of commonly reported group differences leads to concerns that they may do poorly on the test and thus confirm the stereotype. This concern detracts from their ability to focus all of their attention on the test, resulting in poorer test performance. Steele hypothesized a similar effect for gender in the domain of mathematics. A boundary condition for the theory is that individuals must identify with the domain in question. If the domain is not relevant to the individual’s self-image, the testing situation will not elicit stereotype threat.

Steele and Aronson (1995) found support for the theory in a series of laboratory experiments. The basic paradigm used was to induce stereotype threat in a sample of high-achieving majority and minority students statistically equated in terms of their prior performance on the SAT. One mechanism for inducing threat is via instructional set. In the stereotype threat condition, participants were told that they would be given a test of intelligence; in the nonthreat condition, they were told they would be given a problem-solving task. In fact, all of the participants received the same test. Steele and Aronson found a larger majority–minority difference in the threat condition than in the nonthreat condition, a finding supportive of the idea
that the presence of stereotype threat inhibits minority group performance.

These findings are well replicated (Steele, 1997) but commonly misinterpreted. For example, in the fall of 1999, the PBS show “Frontline” broadcast a one-hour special entitled “Secrets of the SAT,” in which Steele’s research was featured. The program’s narrator noted the large Black–White gap on standardized tests, described the stereotype threat manipulation, and concluded, “Blacks who believed the test was merely a research tool did the same as Whites. But Blacks who believed the test measured their abilities did half as well.” The critical fact excluded was that whereas a large score gap exists in the population in general, Steele studied samples of Black and White students who had been statistically equated on the basis of SAT scores. Thus, rather than eliminating the large score gap, the research actually showed something very different. Absent stereotype threat, the Black–White difference was just what one would expect (i.e., zero), as the two groups were equated.

The research actually showed something very different. The critical fact excluded was that whereas a large score gap exists in the population in general, Steele studied samples of Black and White students who had been statistically equated on the basis of SAT scores. Thus, rather than eliminating the large score gap, the research actually showed something very different. Absent stereotype threat, the Black–White difference was just what one would expect (i.e., zero), as the two groups were equated. The presence of stereotype threat, the Black–White difference was larger than would be expected, given that the two groups were equated.

There are a variety of additional issues that cloud interpretation and application of Steele’s (1997) findings. One critical issue is whether the SAT scores used to equate the Black and White students are themselves influenced by stereotype threat, thus confounding interpretation of study findings. A second issue involves questions as to the populations to which these findings generalize (e.g., Whaley, 1998). The work of Steele and coworkers focused on high-ability college students; Steele (1999) noted that the effect is not replicable in the broader population. A third issue is the conflict between a stereotype threat effect and the large literature cited earlier indicating a lack of predictive bias in test use. If stereotype threat results in observed scores for minority group members that are systematically lower than true scores, one would expect underprediction of minority group performance, an expectation not supported in the predictive bias literature. An additional pragmatic issue is the question of how one might reduce stereotype threat in high-stakes testing settings when the purpose of testing is clear.

These issues aside, Steele’s (1997, 1999) research is important in that it clearly demonstrates that the instructional set under which examinees approach a test can affect test results. However, research has yet to demonstrate whether and to what degree this effect generalizes beyond the laboratory. Thus, we caution against overinterpreting the findings to date, as they do not warrant the conclusion that subgroup differences can be explained in whole or in large part by stereotype threat.

The research on test-taker motivation and instructional sets has been conducted primarily in laboratory settings. The effects observed on subgroup differences are not large. Future research should attempt to replicate these findings in a field context so we may better understand the extent to which group differences can be reduced using this alternative. Given the relatively small effects obtained in controlled environments, it seems doubtful that motivational and social effects will account for much of the subgroup differences observed. Nonetheless, it may make sense for test users to institute mechanisms for enhancing motivation, such as the use of more realistic test stimuli clearly applicable to school or job requirements, for the purpose of motivating all examinees.

**Use of Portfolios, Accomplishment Records, and Performance Assessments**

Researchers have experimented with methods that directly measure an individual’s ability to perform aspects of the job or educational domain of interest as a fifth alternative to using paper-and-pencil measures of knowledge, skill, ability, and achievement. Portfolios, accomplishment records, and performance assessments have each been investigated as potential alternatives to traditional tests. Performance assessments (sometimes referred to in the employment domain as job or work samples) require an examinee to complete a set of tasks that sample the performance domain of interest. The intent is to obtain and then evaluate a realistic performance sample in an environment that closely simulates the work or educational setting in which performance assessments may be comprehensive and broad-based, designed to obtain a wide ranging behavior sample reflecting many aspects of the performance domain in question or narrow with the intent of sampling a single aspect of the domain in question. Accomplishment records and portfolios differ from performance assessments in that they require examinees to recount past endeavors or produce work products illustrative of an examinee’s ability to perform across a variety of contexts. Often examinees provide examples demonstrative of their progress toward skill mastery and knowledge acquisition.

Performance assessments, as a potential solution for resolving subgroup differences, were examined in the employment domain by Schmidt, Greenthal, Hunter, Berner, and Seaton (1977), who reported that performance assessments corresponded to substantially smaller Black–White subgroup differences when compared with a written trades test \(d = 0.81 \text{ vs. } 1.44\). N. Schmitt et al. (1996) updated this estimate to \(d = 0.38\) on the basis of a meta-analytic review of the literature, although they combined tests of job knowledge and job samples in their review. The use of performance assessments in the context of reducing subgroup differences has been extended to multiple high-stakes situations in the credentialing, educational, and employment arena. We outline three such efforts here.

**Legal skills assessment center.** Klein and Bolus (1982; described in Sackett, 1998) examined an assessment center developed as a potential alternative to the traditional bar examination. Each day of the two-day center involved a separate trial, with a candidate representing the plaintiff on one day and the defendant on the second. The center consisted of 11 exercises, such as conducting a client interview, delivering an opening argument, conducting a cross-examination, and preparing a settlement plan. Exercises were scored by trained attorneys. Sackett reported a Black–White \(d = 0.76\) and an internal consis-
tency reliability estimate of .67, resulting in a \( d \) corrected for unreliability of 0.93.

**Accomplished teacher assessment.** Jaeger (1996a, 1996b) examined a complex performance assessment process developed to identify and certify highly accomplished teachers under the auspices of the National Board for Professional Teaching Standards. Different assessment packages are developed for different teaching specialty areas. Jaeger examined assessments for Early Childhood Generalists and Middle Childhood Generalists. The assessment process required candidates to complete an assessment center and prepare in advance a portfolio that included videotaped samples of their performance. From a frequency count of scores, we computed Black–White \( ds \) of 1.06 and 0.97 for preoperational field trials, and 0.88 and 1.24 for operational use of the Early and Middle Childhood assessments, respectively.

**Management assessment center.** Goldstein, Yusko, Braverman, Smith, and Chung (1998) examined an assessment center designed for management development purposes. Candidate performance on multiple dimensions was evaluated by trained assessors who observed performance on seven exercises, including an in-basket, leaderless group discussions, and a one-on-one role play. Scores on each exercise corresponded to Black–White \( ds \) ranging from 0.03 to 0.40 for the individual exercises, and a \( d \) of 0.40 for a composite across all exercises.

One clear message emerging from these studies is that it is not the case that an assessment involving a complex realistic performance sample can always be expected to result in a smaller subgroup difference than differences commonly observed with traditional tests. The legal skills assessment center and the accomplished teacher assessment produced subgroup differences comparable in magnitude with those commonly reported for traditional knowledge, ability, and achievement tests. The management assessment center did result in smaller subgroup differences, a finding commonly reported for this type of performance assessment (Thornton & Byham, 1982). We offer several observations as to what might account for the differences in findings across the three performance assessments. First, in the legal skills and accomplished teacher assessments, the assessment exercises required performances that build on an existing declarative knowledge base developed in part through formal instruction. Effectiveness in delivering an opening argument, for example, builds on a foundation of legal knowledge and reasoning. In contrast, typical practice in management assessment centers is to design exercises that do not build on a formal knowledge base, but rather are designed to tap characteristics such as communication skills, organizing and planning skills, initiative, effectiveness under stress, and personal adjustment. Not only do these not build on a formal knowledge base, but they also in many cases reflect dimensions outside the cognitive domain. As a reflection of this difference, we note that the legal skills assessment center correlates .72 with the MBE, whereas the seven exercises in the management assessment center produce correlations corrected for unreliability ranging from .00 to .31 with a written cognitive ability measure.

Thus, although each of these performance assessments reflected highly realistic samples or simulations of the behavior domain in question, the reductions in subgroup differences were, we posit, a function of the degree to which the assessment broadened the set of constructs assessed to include characteristics relevant to the job or educational setting in question, but having little or no relationship with the focal constructs tapped by the traditional test.

Researchers and practitioners in the educational arena have been particularly vocal about the need to evaluate student ability and teacher ability using more than standardized, multiple-choice tests (e.g., D'Costa, 1993; Dwyer & Ramsey, 1995; Harmon, 1991; Lee, 1999; Neil, 1995). Whether termed *authentic assessments* or *constructed-response tests*, these performance assessments have many advocates in and out of the educational community. Until recently, very few attempts had been made to determine whether their use decreases the difference between the measured progress of minority and majority students. Given the high degree of fidelity to actual performance, many consider this alternative a worthwhile approach.

Under particular focus in the education literature is the use of traditional, multiple-choice tests to measure writing skill. It is commonly argued that a constructed-response format that requires examinees to write an essay would serve as a more appropriate measure of writing ability. To that end, a number of studies have compared subgroup differences on multiple-choice tests of writing ability with differences observed on essay tests. Welch, Doolittle, and McLarty (1989) reported that Black–White differences on the traditional ACT writing skills test and an ACT essay test were the same, with \( d = 1.41 \) in both cases. Bond (1995) reported that differences between Blacks and Whites on the extended-essay portion of the National Assessment of Educational Progress (NAEP) were actually greater after correcting for unreliability than those found on the multiple-choice reading portion. In contrast, White and Thomas (1981) reported a decrease in subgroup differences when writing ability was assessed using an essay test versus a multiple-choice test. Using the descriptive statistics reported, we established that the Black–White \( d \) decreased from 1.39 for the traditional test to 0.81 for the essay test, with similar decreases observed for Hispanics and Asians. White and Thomas did not present reliability information, making it impossible to correct these differences for unreliability. However, other studies permitting this correction have reported means and standard deviations that, when translated into \( ds \) corrected for unreliability, suggest similar subgroup differences on essay tests for Blacks and Hispanics (Applebee, Langer, Jenkins, Mullis, & Foertsch, 1990; Koenig & Mitchell, 1988).

Klein et al. (1997) compared subgroup mean differences between racial and ethnic groups on science performance assessments and on the Iowa Tests of Basic Skills, a traditional multiple-choice test. The examinees were fifth, sixth, and ninth graders participating in a statewide, performance assessment effort in California. A high level of score reliability was achieved with the use of multiple
The accomplishment record (Hough, Keyes, & Dunnette, 1983) was developed in part to surmount some of the development and administration cost issues characteristic of performance assessments. Accomplishment records ask examinees to describe major past accomplishments that are illustrative of competence on multiple performance dimensions. These accomplishments are then scored using behaviorally defined scales. Accomplishment records can be used to assess competence in a variety of work and nonwork contexts. We do note that a very similar approach was developed by Schmidt et al. (1979) under the label of the "behavioral consistency method"; a meta-analysis by McDaniel, Schmidt, and Hunter (1988) reported useful levels of predictive validity across 15 studies using this approach.

Hough et al. (1983) used accomplishment records to evaluate attorneys and validated these instruments against performance ratings. The accomplishment records were scored with a high degree of interrater reliability (.75 to .85 across the different performance dimensions and the total score). These scores were then correlated with attorney experience (average r = .24), in order to partial out experience from the relationship between accomplishment record scores and performance ratings. These partialed validity coefficients ranged from .17 to .25 across the dimensions. Validities for a small group of minority attorneys were larger than those for the majority group. Hough (1984), describing the same data, reported Black–White subgroup differences of d = 0.33 for the accomplishment records. The performance ratings exhibited almost exactly the same difference (i.e., d = 0.35). It is interesting that the accomplishment records correlated near zero with the LSAT, scores on the bar exam, and grades in law school. These more traditional measures of ability would most likely have exhibited greater d when compared with the accomplishment records, although Hough did not present the relevant subgroup means and standard deviations. Because the accomplishment records likely measured constructs in addition to ability (e.g., motivation and personality), it is perhaps not surprising that d was lower than that found for more traditional cognitively oriented tests.

Similar to accomplishment records, portfolios represent examinees' past achievements through a collection of work samples indicative of one's progress and ability. Although it is applicable to adult assessment, much of the research involving portfolios has taken place in schools. LeMahieu, Gitomer, and Eresh (1995) reported on a project in the Pittsburgh schools in which portfolios were used to assess students' writing ability in Grades 6–12. Their experience indicated that portfolios could be rated with substantial interrater agreement. They also reported that Black examinees' scores were significantly lower than those of White examinees, but did not report subgroup means precluding the estimation of an effect size. Supovitz and Brennan (1997) reported on an analysis of writing portfolios assembled by first and second graders in the Rochester, New York, schools. Scores on two standardized tests were compared with scores based on their portfolios. Interrater reliability of the scoring of the language arts portfolios was .73 and .78 for first and second graders, respectively, whereas the reliability of the two standardized tests was .92 and .91. Differences between Black and White students were about twice as large on the standardized tests as they were on the writing samples. On both tests, the differences between subgroups were smaller (0.25 to 0.50 in standard deviation units depending on the test type) than is usually reported.

Although accomplishment records and portfolios likely have fewer development costs when compared with performance assessments, the cost of scoring, especially if multiple raters are used, may be high. Another issue present with accomplishment records specifically is the reliance on self-report, although an attempt to secure verification of the role of the examinee in each accomplishment may diminish the tendency to overreport or enhance one's role. There may also be differing levels of opportunity to engage in activities appropriate for portfolios or accomplishment records. To the extent that examinees feel that they do not have the resources available to assemble these extensive documents, they may find the experience demotivating and frustrating. This concern is important inasmuch as there is some evidence (Ryan, Ployhart, Greguras, & Schmit, 1998; Schmit & Ryan, 1997) that a greater proportion of minority than majority individuals withdraw during the various hurdles in a selection system.

Summary. Use of more realistic or authentic assessments does not eliminate or even diminish subgroup
differences in many of the educational studies. Also, all of the studies report that the reliable scoring of these tests is difficult and expensive to achieve in any large-scale testing application. Problems with the standardization of the material placed in portfolios and the directions and opportunities afforded students are also cited in studies of student reactions to the use of these tests (Dutt-Doner & Gilman, 1998) as well as by professionals. Accomplishment records and job samples used in employment contexts show smaller subgroup differences in some studies than do cognitively loaded tests. The attribution that these smaller subgroup differences are due to test type are probably unwarranted, as scores on most job samples and accomplishment records most likely reflect a mix of constructs that go beyond those measured by traditional knowledge, skill, ability, and achievement tests.

Use of Coaching or Orientation Programs

Another strategy for reducing subgroup differences is the use of coaching or orientation programs. The purpose of these programs is to inform examinees about test content, provide study materials, and recommend test-taking strategies, with the ultimate goal of enabling optimal examinee performance. The term coaching is at times used to refer to both orientation programs that focus on general test-taking strategies and programs featuring intensive drill on sample test items. We use the term orientation programs to refer to short duration programs, dealing with broad test-taking strategies, that introduce examinees to the types of items they will encounter. We use the term coaching to refer to more extensive programs, commonly involving practice and feedback, in addition to the material included in orientation programs. A review by Sackett, Burris, and Ryan (1989) indicated that coaching programs involving drill and practice do show evidence of modest score gains above those expected due simply to retesting. Although there is little literature on the differential effectiveness of coaching and orientation programs by subgroup, a plausible hypothesis is that subgroups differ in their familiarity with test content and test-taking skills. This difference in familiarity may contribute to observed subgroup differences in test scores. Conceivably, coaching or orientation programs would reduce error variance in test scores due to test anxiety, unfamiliar test formats, and poor test-taking skills (Frierson, 1986; Ryan et al., 1998) that would in turn reduce the extent of subgroup differences. However, there is evidence suggesting the presence of a larger coaching effect for individuals with higher precoaching test scores, a finding that argues against the likelihood that coaching will narrow the gap between a lower scoring subgroup and a higher scoring subgroup. With that caveat, we discuss below the coaching and orientation literature investigating the influence of this strategy on subgroup differences.

Ryan et al. (1998) studied an optimal orientation program that familiarized firefighter job applicants with test format and types of test questions. The findings indicated that Blacks, women, and more anxious examinees were more likely to attend the orientation sessions, but attending the orientation program was unrelated to test performance or motivation. Ryer, Schmidt, and Schmitt (1999) studied a mandatory orientation program for entry-level jobs in a manufacturing organization at two locations, with each location having a control group and a test orientation group. The results showed a small positive impact of orientation on the test scores of minority examinees, approximately 0.15 in standard deviation units, and the applicants did indicate that they view organizations that provide these programs favorably. However, the orientation program had greater benefits for nonminority members than minority members at one of the two locations. Schmit (1994) studied a voluntary orientation program for police officers that consisted of a review of test items and content, recommendations for test-taking strategies, practice on sample test items, and suggestions on material to study. Attendance at the program was unrelated to race, and whereas everyone who attended the program scored higher on the examination than did nonattenders, Black gains were twice as large as were those of Whites. No standard deviation was provided for the test performance variable, so d could not be estimated.

The educational literature includes a relatively large number of efforts to evaluate coaching initiatives. At least three reviews of this literature have been conducted. Mesick and Jungeblut (1981) reported that the average difference between coached examinees and noncoached examinees taking the SAT was about 0.15 standard deviation units. The length of the coaching program and the amount of score gains realized were positively correlated. Mesick and Jungeblut estimated that a gain of close to 0.25 standard deviation units could be achieved with a program that would approach regular schooling. DerSimonian and Laird (1983) reported an average effect size of 0.10 standard deviation units for coaching programs directed at the SAT test, an aptitude test. In an analysis of coaching programs directed at achievement tests, Bangert-Downs, Kulik, and Kulik (1983) reported gains of about 0.25 standard deviation units as a function of coaching. Thus, the effects of coaching on performance on traditional paper-and-pencil tests of aptitude and achievement appear to be small, but replicable.

Frierson (1986) outlined results from a series of four studies investigating the effects of test-taking interventions designed to enhance minority examinee test performance on various standardized medical examinations (e.g., Medical College Admissions Test [MCAT], Nursing State Board Examination). The programs taught examinees test-taking strategies and facilitated the formation of learning-support groups. Those minorities who experienced the interventions showed increased test scores. However, the samples used in these studies included very few White examinees, making it difficult to discern whether coaching produced a differential effect on test scores in favor of minorities. Powers (1987) reexamined data from a study on the effects of test preparation involving practice, feedback on results, and test-taking strategies using the initial version of the GRE analytical ability test (Powers & Swinton, 1982, 1984). The findings indicated that when supplied with the same test preparation materials, no particular
subgroup appeared to gain more or less than any other subgroup, although all of the examinees showed statistically significant score gains. Koenig and Leger (1997) evaluated the impact of test preparation activities undertaken by individuals who had previously taken the MCAT but had not passed. Black examinee scores improved across two administrations of the MCAT less than did those of White examinees.

Overall, the majority of studies on coaching and orientation programs indicate that these programs have little positive impact on the size of subgroup differences. These programs do benefit minority and nonminority examinees slightly, but they do not appear to reduce subgroup differences. Note that these programs would affect passing rates in settings such as licensure, in which performance relative to a standard is at issue, rather than performance relative to other examinees. On the positive side, these programs are well received by examinees; they report favorable impressions of those institutions that offer these types of programs. Additional questions also of interest focus on examinee reactions to these programs, the effectiveness of various types of programs, and ways to make these programs readily available to all examinees.

**Use of More Generous Time Limits**

A final option available for addressing group differences in test scores is the strategy of increasing the amount of time allotted to complete a test. Unless speed of work is part of the construct in question, it can be argued that time limits may bias test scores. Tests that limit administration time may be biased against minority groups, in that certain groups may be provided too little time to complete the test. It has been reported that attitudes toward speededness are culture-bound with observed differences by race and ethnicity (O’Connor, 1989). This suggests that providing examinees with more time to complete a test may facilitate minority test performance.

Research to date, however, is not supportive of the notion that relaxed time limits reduce subgroup differences. Evans and Reilly (1973) increased the time that examinees were allotted to complete the Admission Test for Graduate Study in Business from 69 seconds/item to 80 seconds/item and 96 seconds/item. The result was a corresponding increase in subgroup differences from a Black–White $d$ of 0.83 to Black–White $d$s of 1.12 and 1.38, respectively. Wild, Durso, and Rubin (1982) investigated whether reducing the speededness of the GRE verbal and quantitative tests from 20 minutes to 30 minutes reduced subgroup differences. Results using both operational and experimental versions of the tests indicated that whereas increasing the time allotted benefited all examinees, it did not produce differential score gains favoring minorities and often exacerbated the extent of subgroup differences already present. Applebee et al. (1990) reported that doubling the time that 4th, 8th, and 12th graders were given when completing NAEP essay tests led to inconsistent changes in the subgroup differences observed for Blacks and Hispanics. Through the use of odds ratios computed on the basis of the percentage of examinees that received an adequate or better passing score, we ascertained that the Black–White $d$ increased on six of the nine essay tests and the Hispanic–White $d$ increased on five of the nine essay tests when the tests were administered with extended time. Thus, the time extension increased the differences between Black and White essay scores and Hispanic and White essay scores in the majority of test administrations.

Although relaxed time limits will likely result in higher test scores for all examinees, there does not appear to be a differential benefit favoring minority subgroups. In fact, it is more common that extending the time provided to examinees when completing a test increases subgroup differences, sometimes substantially. Time limits enhance the error in test scores resulting from differences in pace of work versus quality of work. Interpreting the observed increases in subgroup differences when these limits are removed suggests that the changes observed are likely a function of enhancing the reliability of the test. Thus, time extensions are unlikely to serve as an effective strategy for reducing subgroup differences.

**Discussion and Conclusions**

Our goal in this article was to consider a variety of strategies that have been proposed as potential methods for reducing the subgroup differences regularly observed on traditional tests of knowledge, skill, ability, and achievement. Such tests are commonly used in the contexts of selection for employment, educational admissions, and licensure and certification. We posit that there is extensive evidence supporting the validity of well-developed traditional tests for their intended purposes, and that institutions relying on traditional tests value the positive outcomes resulting from test use. Thus, one set of critical constraints is that any proposed alternative must include the constructs underlying traditional tests and any proposed alternative cannot result in an appreciable decrement in validity. Put another way, we focus here on the situation in which the institution is not willing to sacrifice validity in the interest of reducing subgroup differences, as it is in that situation that there is tension between pursuing a validity-maximization strategy and a diversity-maximization strategy. A second set of critical constraints is that, in the face of growing legal obstacles to preference-based forms of affirmative action, any proposed alternative cannot include any form of preferential treatment by subgroup. Within these constraints, we considered two general types of strategies: modifications to procedural aspects of testing and the creation of alternative testing instruments.

In evaluating the various strategies that have been promulgated as potential solutions to the performance versus diversity dilemma, our review echoes the sentiments of the 1982 National Academy of Sciences panel that concluded, "the Committee has seen no evidence of alternatives to testing that are equally informative, equally adequate technically, and also economically and politically viable" (p. 144). Although our observations are offered almost 20 years later, the story remains much the same. Alternatives to traditional tests tend to produce equivalent subgroup differences in test scores when the alternative test
measures cognitively loaded constructs. If such differences are not observed, the reduction can often be traced to an alternative that exhibits low levels of reliability or introduces noncognitive constructs. In fact, certainly the most definitive conclusion one can reach from this review is that adverse impact is unlikely to be eliminated as long as one assesses domain-relevant constructs that are cognitively loaded. This conclusion is no surprise to anyone who has read the literature in this area over the past three or more decades. Subgroup differences on cognitively loaded tests of knowledge, skill, ability, and achievement simply document persistent inequities. Complicating matters further, attempts to overcome issues associated with reliable measurement often result in a testing procedure that is cost-prohibitive when conducted on a large scale. In spite of these statements, there are a number of actions that can be taken by employers, academic admissions officers, or other decision makers who are faced with the conflict between diversity goals and a demand that only those who are most able should be given desirable educational and employment opportunities. Although elimination of subgroup differences via methods reviewed in this article is not feasible, reduction in subgroup differences, if it can be achieved without loss of validity, would be of considerable value.

First, in constructing test batteries, the full range of performance goals and organizational interests should be considered. In the employment arena, researchers have tended to focus on measures of maximum performance (i.e., ability), rather than on measures of typical performance (perhaps most related to motivation factors), when considering what knowledge, skills and abilities to measure. These maximum performance constructs were easy to measure using highly reliable and valid instruments. With the recent literature espousing the value of personality, improvements in interviews, and better methods for documenting job-related experience, valid methods for measuring less cognitively oriented constructs are becoming available. When these constructs are included in test batteries, there is often less adverse impact. We must also emphasize the importance of clearly identifying the performance construct one is hoping to predict. The weighting of different aspects of performance and organizational goals should determine the nature of the constructs measured in a high-stakes testing situation. It is important to measure what is relevant, not what is convenient, easy, or cheap.

Second, research on the identification and removal of items that may be unfairly biased against one group or another does not indicate that any practically significant reductions in $d$ can be achieved in this fashion. Studies of DIF are characterized by small effects, with items not consistently favoring one group versus another. The effects of removing biased items on overall test characteristics are usually minimal. It does seem apparent that one should write as simply as possible consistent with the construct one is hoping to measure and that content that is obviously cultural should be removed.

Research on the mode of presenting test stimuli suggests that video-based procedures, which broaden the range of constructs assessed, or reducing the verbal component (or reading level) of tests may have a positive effect on subgroup differences, although $d$ is often still large enough to produce adverse impact, particularly when the selection ratio is low. Results are not consistent across studies, and clearly more research would be helpful. Such studies are particularly difficult to conduct. Separation of the mode of testing and the construct tested is a challenge; conflicting results across studies may be due to an inability to differentiate between constructs and methods. With improvements in technology, alternatives to traditional paper-and-pencil tests are clearly feasible and worthy of exploration. It is also important to note that verbal ability may be a skill related to important outcomes and hence considered a desirable component of test performance. In these cases, it would be best to include a measure that specifically assesses verbal ability so that one may remove its influence when measuring other job-related constructs. The entire test battery can then be constructed to reflect an appropriate weighting and combination of relevant attributes given the relative importance of the various performance outcomes.

Whenever possible, it seems desirable to measure experiences that reflect necessary knowledge, skills, and abilities required in the target situation. Accomplishment records reduced differences compared with the usual $d$ obtained with cognitively loaded tests in a manner that was practically important as well. This is very likely because additional constructs are targeted and assessed in the accomplishment record. Results for portfolio and performance assessments in the educational arena have been mixed. Some studies indicate lower levels of $d$, whereas other studies indicate no difference or even greater differences on portfolio or performance assessments when compared with the typical multiple-choice measure of achievement. Differences in the level of $d$ across studies may be due partly to the degree to which test scores are a function of ability and motivation. If partly a function of motivation, we would expect $d$ to be smaller. Again, if relatively complex and realistic performance assessments involve cognitive skills as opposed to interpersonal skills, the level of $d$ will likely be the same as a traditional cognitively loaded measure. In addition, problems in attaining reliable scores at reasonable expense questions the feasibility of this strategy.

It seems reasonable to recommend that some form of test preparation or orientation course be provided to examinees. The effects of coaching appear to be minimally positive over all groups, even though coaching does not seem to reduce $d$. Reactions to test preparation and coaching efforts among job applicants have been universally positive. Insofar as some candidates do not have access to informal networks that provide information on the nature of exams, these programs could serve to place all examinees on the same playing field. At the very least, it would seem that such positive reactions would lead to less complaints about the test and probably less litigation—even though we have little research documenting the relationship between reactions and organizational outcomes.

Finally, we recommend that test constructors pay attention to face validity. When tests look appropriate for the
performance situation in which examinees will be expected to perform, they tend to react positively. Such positive reactions seem to produce a small reduction in the size of $d$. Equally important, perhaps, may be the perception that one’s attributes are being fairly and appropriately assessed.

We note that in this article we have focused on describing subgroup differences resulting from different measurement approaches. We cannot in the space available here address crucial questions of interventions to remedy subgroup differences in the life opportunities that affect the development of the knowledge, skill, ability, and achievement domains that are the focus of this article. The research discussed in this article, suggesting that subgroup differences are not simply artifacts of paper-and-pencil testing technologies, highlights the need to consider those larger questions.

REFERENCES

Hopwood v. State of Texas, 78 F. 3d 932, 948 (5th Cir. 1996).
Jaeger, R. M. (1996b). Conclusions on the technical measurement quality...


