# AN EXAMINATION OF FACTORS CONTRIBUTING TO A REDUCTION IN RACE–BASED SUBGROUP DIFFERENCES ON A CONSTRUCTED RESPONSE PAPER–AND–PENCIL TEST OF ACHIEVEMENT

A Dissertation

by

BRYAN D. EDWARDS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2003

Major Subject: Psychology

# AN EXAMINATION OF FACTORS CONTRIBUTING TO A REDUCTION IN RACE–BASED SUBGROUP DIFFERENCES ON A CONSTRUCTED RESPONSE PAPER–AND–PENCIL TEST OF ACHIEVEMENT

A Dissertation

by

BRYAN D. EDWARDS

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

| | |
|---|---|
| Winfred Arthur, Jr.<br>(Chair of Committee) | Robert D. Pritchard<br>(Member) |
| John F. Finch<br>(Member) | Steve Rholes<br>(Head of Department) |
| Victor Willson<br>(Member) | |

August 2003

Major Subject: Psychology

# ABSTRACT

An Examination of Factors Contributing to a Reduction in Race–Based Subgroup Differences on a Constructed Response Paper–and–Pencil Test of Achievement.

(August 2003)

Bryan D. Edwards, B.S., The University of Alabama;

M.S., University of South Alabama

Chair of Advisory Committee: Dr. Winfred Arthur, Jr.

The objectives of the present study were to: (a) replicate the results of Arthur et al. (2002) by comparing race–based subgroup differences on a multiple–choice and constructed response test in a laboratory setting using a larger sample, (b) extend their work by investigating the role of reading ability, test–taking skills, and test perceptions that could explain why subgroup differences are reduced when the test format is changed from multiple–choice to a constructed response format, and (c) assess the criterion–related validity of the constructed response test. Two hundred sixty White and 204 African Americans completed a demographic questionnaire, Test Attitudes and Perceptions Survey, a multiple–choice or constructed response test, the Raven's Advanced Progressive Matrices Short Form, the Nelson–Denny Reading Test, Experimental Test of Testwiseness, and a post–test questionnaire. In general, the *pattern* of results supported the hypotheses in the predicted direction. For example, although there was a reduction in subgroup differences in performance on the constructed response compared to the multiple–choice test, the difference was not

statistically significant. However, analyses by specific test content yielded a significant reduction in subgroup differences on the science reasoning section. In addition, all of the hypothesized study variables, with the exception of face validity, were significantly related to test performance. Significant subgroup differences were also obtained for all study variables except for belief in tests and stereotype threat. The results also indicate that reading ability, test–taking skills, and perceived fairness partially mediated the relationship between race and test performance. Finally, the criterion–related validity for the constructed response test was stronger than that for the multiple–choice test. The results suggested that the constructed response test format investigated in the present study may be a viable alternative to the traditional multiple–choice format in high–stakes testing to solve the organizational dilemma of using the most valid predictors of job performance and simultaneously reducing subgroup differences and subsequent adverse impact on tests of knowledge, skill, ability, and achievement. However, additional research is needed to further demonstrate the appropriateness of the constructed response format as an alternative to traditional testing methods.

# ACKNOWLEDGMENTS

First, a special acknowledgment goes to my advisor, Winfred Arthur, Jr. for his help and guidance throughout this entire process. My level of success in graduate school would not have been attainable without his gracious mentoring, expertise, patience, support, and trust.

I would like to thank my committee members, Bob Pritchard, Victor Willson, and John Finch for their helpful insights and suggestions for improving upon the research design and with the analyses for this project. I would also like to thank Dr. James Campion, Dr. Peter Metofe, and Kelly Bolton for their help with participant recruitment and data collection.

I am also indebted to my family, including my new family, for their love and encouragement. My mother and father sowed the seeds of success early on in life and taught me to keep all things in perspective.

Finally, I would like to dedicate this dissertation to my wife, Claire, whose patience and self–sacrifice has been instrumental for its completion. I have been truly blessed by her show of love, support, and encouragement. She has been an extraordinary influence in my life and as such, is directly responsible for my personal and professional development. During the last four years, Claire has brought me an indescribable happiness and I am ecstatic about what lies ahead as our new adventure begins.

TABLE OF CONTENTS

Page

# LIST OF FIGURES

FIGURE

# LIST OF TABLES

TABLE

# INTRODUCTION

Arthur, Edwards, and Barrett (2002) found evidence that African American–White test score differences on a cognitively loaded knowledge test were much lower using a constructed response than a multiple–choice format. However, given the operational nature of the data, they were unable to empirically assess reasons for this reduction in subgroup differences. Therefore, the objectives of the present study were to: (a) replicate the results of Arthur et al. (2002), (b) extend their work by investigating the role of factors that could explain why race–based subgroup differences may be reduced when the test format is changed from multiple–choice to constructed response format, and (c) assess the criterion–related validity of the constructed response test. These objectives were accomplished using a college student sample in an academic setting.

In the testing and personnel selection literature, paper–and–pencil tests of ability and achievement have been shown to be the most valid predictors of job and training performance (Schmidt & Hunter, 1998). However, it has also been extensively documented that cognitively loaded paper–and–pencil tests of knowledge, skill, ability, and achievement generally display large subgroup differences (Bobko, Roth, & Potosky, 1999; Hartigan & Wigdor, 1989) with a widely cited one standard deviation difference in African American–White performance (Chan & Schmitt, 1997; Roth, Bevier, Bobko, Switzer, & Tyler, 2001; Schmitt, Clause, & Pulakos, 1996).

This one standard deviation African American–White difference on paper–and–pencil tests of ability and achievement typically results in adverse impact

---

This dissertation follows the style and format of *Journal of Applied Psychology*.

against African Americans. A test or assessment tool displays adverse impact if there are differential outcomes associated with the use of the test (e.g., selection, promotion) as a function of a protected class status variable (i.e., race, sex, color, religion, national origin, age, and disability). The Uniform Guidelines (Equal Employment Opportunity Commission, 1978) established the 80% or the four–fifths rule as a general rule of thumb for operationalizing adverse impact. Using this rule, adverse impact on a selection instrument occurs against a specified subgroup of a protected class when the proportion of the subgroup class members meeting a cut score on the instrument is less than 80% of the proportion of other subgroup class members meeting the same cut score. Other operationalizations of adverse impact include statistical significance tests such as Fisher's exact test.

In the context of testing and personnel selection, race has been the primary, although not only, Title VII protected class of interest. The presence of subgroup differences and associated adverse impact has important implications for individuals, organizations, and society at large. For individuals adversely impacted by a test, there is a loss of employment opportunities and all the benefits associated with such. There are also psychological and emotional costs that often result from being adversely impacted by a selection test. From the organization's perspective, there is the dilemma of using the most valid predictors of job performance and concurrently minimizing the legal, ethical, and professional liabilities associated with using tests that display adverse impact (Sackett, Schmitt, Ellingson, & Kabin, 2001). Thus, subgroup differences present a special dilemma for organizations which simultaneously seek to reduce

adverse impact on tests and still preserve the advantages associated with using valid predictors of job performance.

At the societal level, there are issues pertaining to addressing past social wrongs, diversity, and equal opportunity (Doverspike, Taylor, & Arthur, 1999). The racial and ethnic composition of society's workforce and hence socioeconomic strata is directly affected by decisions based on high–stakes testing. For example, standardized achievement tests such as the Scholastic Assessment Test (SAT) are used for determining admissions and scholarship allocations for college and graduate schools. Licensing and certification exams and employment tests, which measure knowledge, skills, and abilities also influence a society's workforce (Sackett et al., 2001). Many opponents of standardized preemployment selection systems, especially those that yield adverse impact against protected classes, argue that minority group membership warrants preferential treatment in the selection process to compensate for the long history of discrimination and social inequities suffered by members of some minority groups. This position is predicated on the assumption that these observed subgroup differences are the direct or indirect result of past legal discrimination (for reviews of this issue see Crosby, Iyer, Clayton, & Downing, 2003; Doverspike & Arthur, 1995; Doverspike et al., 1999; Taylor–Carter, Doverspike, & Cook, 1995). Therefore, adverse impact poses a stumbling block for equal opportunity employers and organizations with affirmative action programs that value increasing workplace diversity to compensate for past inequities. In some instances, racial diversity is encouraged by legislation and mandated by the courts for organizations that have historically discriminated against minority employees. Previously successful attempts

to increase workplace diversity, such as subgroup norming and affirmative action have been legally challenged or prohibited (Civil Rights Act, 1991; Hopwood v. State of Texas, 1996). Thus, finding other means of reducing adverse impact has unquestionable value.

However, one thing that remains unclear is whether the widely cited one standard deviation difference on cognitively loaded paper–and–pencil tests of knowledge, skill, ability, and achievement is a construct or method phenomenon (Arthur, Day, McNelly, & Edens, 2003; Arthur et al., 2002; Roth, Bevier et al., 2001; Schmitt et al., 1996). That is, do test score differences reflect real race differences on the underlying construct (e.g., cognitive ability) or do they reflect test bias such that the observed subgroup differences are artifactual and not real differences on the underlying construct?

Environmental (Ceci & Williams, 1997; Fischer, Houte, Jankowski, Lucas, Swidler, & Voss, 1996) and biological and genetic (Jensen, 1985; Rushton & Ankey, 1996) explanations have been offered for observed subgroup differences on tests of knowledge, skill, ability and achievement. For example, large correlations (e.g., .50 – .90) are commonly reported for the relationship between cognitive ability and education level (Ceci, 1991; Herrnstein & Murray, 1994; Neisser et al., 1996). Most theories describe this relationship as causal with higher levels of cognitive ability resulting in higher amounts of education. However, Ceci (1991) and Ceci and Williams (1997) argue that the reverse causal relationship may be just as strong and review evidence for the environmental determinants of cognitive ability. Specifically, they provide evidence for the hypothesis that education level significantly affects cognitive ability

(i.e., IQ scores) beyond the more widely accepted causal relationship—that cognitive ability affects educational/academic performance. Therefore, to the extent that there are race–based subgroup differences in education level, these differences in education may be manifested in measures of cognitive ability.

In terms of genetic explanations, Jensen (1985) supports Spearman's hypothesis that African American–White test score differences are more pronounced on cognitive ability (*g*) tests with speed of information processing postulated to be the mechanism that results in the observed subgroup differences. Furthermore, Rushton (2000) espouses genetic explanations for racial test score differences by citing evidence for smaller brain size, and consequently, slower information processing speed in African Americans than Whites. Rushton and Ankey (1996) present a qualitative review of the literature on the relationship of brain size and cognitive ability in terms of race, age, sex, and socioeconomic status variables. The authors review research studies that suggest the existence of a "race gradient" in which cognitive ability and brain size is greater in East Asians than Europeans which in turn is greater than Africans. The authors provide several theories that would explain race differences in cognitive ability and brain size; however, they do not present any empirical evidence that directly support any of these theories. Furthermore, these theories of subgroup differences have been generally discounted and are not widely accepted (Helms, 1992; Zuckerman, 1990).

A third explanation for subgroup differences found in the personnel selection and testing literature has focused on test bias (Arthur & Doverspike, 2003; Reynolds & Brown, 1984; Sackett et al., 2001; Schmitt, 1989). Specifically, this view posits that

racial test score differences are artifactual and not real. Sources of artifactual variance include: the effects of test–taking skills that are not job relevant (Ryan & Greguras, 1998), the use of culture–specific language on the specified tests (Helms, 1992), and the use of test formats (e.g., multiple–choice) that discriminate among subgroups via processes of test perceptions or reading demands that are not job–related (Chan & Schmitt, 1997; Ryan & Greguras, 1998).

Prior to 1991, a common practice was to adjust scores to produce similar subgroup distributions thereby eliminating test bias. The reasoning behind this strategy was that scores on specified tests represent different constructs for different subgroup classifications. Thus, a mean score of 80 on a test for one group was equivalent to an 85 or 90 for another group. Therefore, tests were renormed within subgroups to equate scores across subgroups.

Rank ordering candidates on the basis of predictor raw scores, specifically in the presence of subgroup differences, will yield higher selection rates for the higher scoring racial subgroup (e.g., Whites or males) resulting in adverse impact against the lower scoring subgroup (e.g., African Americans or females). However, subgroup differences can be statistically removed by converting raw test scores to standard scores or percentiles within each subgroup (i.e., subgroup norming) or by adding a constant to the scores of the lower–scoring racial subgroups to eliminate systematic test score differences. The objective is to eliminate the difference in distributions as a function of group membership so that individual candidates can be directly compared on the same underlying distribution of scores (Sackett & Wilk, 1994). Consequently, rank ordering candidates using adjusted scores will equate selection ratios on the predictor, thereby

eliminating adverse impact against specified subgroups of protected classes. However, passage of the Civil Rights Act of 1991 (CRA, 1991) made the practice of any test score adjustment on the basis of protected class variables illegal. Since the ban on subgroup or race norming and other adjustments to test scores on the basis of protected class status (CRA, 1991), attempts to reduce adverse impact have taken one of two approaches—the use of predictor constructs other than knowledge, skill, ability or achievement with lower cognitive demands and the use of alternative testing methods.

### *Changing the Construct*

Common noncognitive constructs that have been investigated in attempts to reduce adverse impact have included personality variables (Hogan, Hogan, & Roberts, 1996) such as integrity (Ones, Viswesvaran, & Schmidt, 1993; Sackett & Wanek, 1996) and conscientiousness (Schmitt et al., 1996). Integrity has been examined as an alternative predictor of job performance and generally displays no subgroup differences. Ones et al.'s (1993) meta–analysis of integrity test validities indicates that the best estimate of the true mean correlation between integrity tests and supervisor ratings of job performance is .41. The relationship between integrity and counterproductive behaviors (e.g., theft, absenteeism) is also quite high (i.e., .47), but moderated by several methodological variables such as type of test (overt vs. personality–based) and sample (applicants vs. incumbents).

Ones et al. (1993) reported large racial subgroup differences for overt integrity tests (.72; nonminority mean was higher than the minority) and somewhat smaller mean differences for personality–based integrity tests (.20). However, several other researchers (e.g., Sackett & Wanek, 1996; Terris & Jones, 1982) have failed to obtain

subgroup differences in integrity test means between African Americans and Whites. For instance, Sackett and Wanek (1996) report that a reanalysis of integrity test mean subgroup differences by Ones, Viswesvaran, and Schmidt (1996) failed to obtain any racial subgroup differences. Sackett and Wanek (1996) note that the large subgroup differences reported by Ones et al. (1993) in their original meta–analysis were due to computational errors. Thus, the extant literature suggests that if integrity were used as the sole predictor of job performance for selection, there would be no adverse impact against African American applicants.

The Big Five personality dimensions are often used as alternatives to knowledge, skill, ability and achievement in the prediction of job performance (Barrick & Mount, 1991; Hurtz & Donovan, 2001). Correlations between the Big Five and overall job performance reported in the literature typically range from .03 – .22 (Hurtz & Donovan, 2001). However, larger validities are reported between *specific components* of each of the Big Five and *specific facets* of the job performance domain. For example, Hough (1992) reported an uncorrected correlation of -.38 between Achievement (a component of the Conscientiousness construct) and counterproductive behaviors.

Most of the empirical evidence indicates that measures of the Big Five personality constructs do not typically yield racial subgroup differences (Hogan et al., 1996; Hough, Oswald, & Ployhart, 2001; Sackett & Wilk, 1994). Hough (1998) computed standardized differences between racial subgroups using normative data obtained from manuals of several commercially available personality tests. African American–White subgroup differences were minimal, ranging from *d*s of .00 to -.31.

The largest subgroup differences were reported for Affiliation ($d$ = -.31) and Intellectance ($d$ = -.28) with African Americans scoring lower. In contrast, African Americans scored higher on Potency ($d$ = .15). Hispanic and White subgroup standardized differences on the same personality scales were less than .10 with the exception of Unlikely Virtues in which Hispanics scored higher ($d$ = .60). Hough et al. (2001) report results similar to those described above.

Furthermore, Collins and Gleaves (1998), using confirmatory factor analysis, found evidence that the five–factor model of personality fit equally well in a sample of White and African American job applicants. This research was in response to criticisms that the Big Five factor structure is quite different for African Americans and Whites—an issue that could have serious implications for the use of these noncognitive predictors in selection contexts (Azibo, 1991; LaFromboise, Coleman, & Gerton, 1993). Finally, one study did report evidence of racial subgroup differences on two personality variables which resulted in adverse impact at low selection ratios (Ryan, Ployhart, & Friedal, 1998). In summary, existing evidence suggests there are minimal to no racial subgroup differences on most job–relevant personality constructs.

In terms of criterion–related validities, Barrick and Mount (1991) found conscientiousness to have the strongest relationship to job performance, with correlations ranging from .20 to .23. In addition, Hurtz and Donovan's (2001) meta–analysis obtained a true estimated correlation of .20 between conscientiousness and overall job performance. That is, of all the Big Five factors, the highest "true" population correlation appears to be around .20, which in terms of practical significance, represents a small amount of explained variance (i.e., 4%) in job

performance. For instance, in terms of Cohen's (1992) rules of thumb (i.e., small = .10, medium = .30, large = .50 and Hemphill's (2003) empirical guidelines (i.e., small = <.20, medium = .20 to .30, large = >.30) for interpreting correlational effect sizes, the strongest population correlation between personality and job performance can best be described as small to medium. In contrast, the predictive validity of other constructs such as knowledge, skill, ability, achievement, and training and experience is much higher (Schmidt & Hunter, 1998). Therefore, although there appears to be an absence of subgroup differences, one could conclude that the criterion–related validity of personality variables may not be strong enough to warrant what at present, appears to be their increasing use in selection contexts and the extant literature. In addition, a number of researchers (e.g., Arthur, Woehr, & Graziano, 2001; Ryan et al., 1998) have raised several overlooked issues in the application of personality variables in personnel selection contexts and provide some compelling reasons to carefully consider these issues when using personality tests in selection and other organizational decision–making contexts.

Other noncognitive constructs that have been proposed are practical intelligence and tacit knowledge (Sternberg & Hedlund, 2002; Sternberg, Wagner, Williams, & Horvarth, 1995). Practical intelligence refers to common sense or "street smarts" as opposed to "book smarts" which is represented by cognitive ability or traditional intelligence. Tacit knowledge is defined as "action–oriented knowledge, acquired without direct help from others, that allows individuals to achieve goals they personally value" (p. 916, Sternberg et al., 1995). Sternberg et al. (1995) review empirical studies that support the validity of these constructs in predicting various outcomes such as job

performance. In addition, they emphasize that these constructs are distinctly different from traditional intelligence (e.g., operationalized as I.Q.) and that the correlation between these two constructs is zero. Finally, they argue that there are no race–based subgroup differences in practical intelligence or tacit knowledge. For example, they cited one study (i.e., Eddy, 1988) that reported a .03 correlation between race and tacit knowledge. Thus, they conclude that this construct is a viable alternative to traditional cognitively loaded constructs for resolving the organizational dilemma.

In summary, many of the noncognitive predictor constructs proposed as alternatives to knowledge, skill, ability and achievement have been found to be related to job performance to some degree; but, with the exception of integrity, the correlations are substantially lower than those for the relationship between cognitively loaded constructs and performance (Schmidt & Hunter, 1998). And, although subgroup differences have been shown to be lower on some of these constructs (e.g., personality; Hogan et al., 1996; Schmitt, et al., 1996), the use of noncognitive predictor constructs in reducing adverse impact has generally not been very successful (Ryan et al., 1998) and the lower validity resulting from the use of these constructs may result in a considerable reduction in utility (Hunter & Hunter, 1984; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997).

### *Combining Constructs*

A variation of the approach of using noncognitive predictor constructs is to *combine* cognitively loaded constructs with other predictors (e.g., personality variables). The intent is to combine the high validity of paper–and–pencil tests of knowledge, skill, ability and achievement with the low subgroup differences often

observed with noncognitive predictors to create a composite selection battery that solves the organizational dilemma of coming up with valid predictors with low levels of adverse impact. Thus, researchers have looked at optimal combinations of predictors (Pulakos & Schmitt, 1996; Sackett & Ellingson, 1997; Sackett & Roth, 1996; Schmitt et al., 1997) and criteria (i.e., different facets of job performance; Hattrup, Rock, & Scalia, 1997) in an attempt to reduce adverse impact and minimize the loss of selection utility.

A few studies that examined the approach of combining cognitively loaded constructs with other predictors documented an enhancement in the validity of the selection battery and simultaneous reductions in adverse impact (Ones et al., 1993; Pulakos & Schmitt, 1996). Pulakos and Schmitt (1996) examined the predictive validity and subgroup differences in a multi–predictor battery that consisted of biodata, a situational judgement test, a structured interview, and a paper–and–pencil and video–based verbal ability test. They found that validity was enhanced with the use of biodata, the situational judgement test, and the structured interview in combination with the verbal ability test and that subgroup differences were minimized. However, the 80% rule was still violated unless the verbal ability test was omitted from the selection battery. Furthermore, structured interviews and situational judgement tests are methods and the authors only implied that they measured noncognitive predictor constructs, without indicating which constructs were measured.

Ones et al. (1993) found evidence for the incremental validity of integrity tests over cognitive ability. Specifically, they estimated that integrity tests, when combined with cognitive ability increased the multiple validity correlation by 22% – 100% over using cognitive ability alone depending on the level of job complexity. Furthermore,

Schmidt and Hunter (1998) reported a 27% increase in the prediction of job performance using integrity in conjunction with cognitive ability. However, Schmidt and Hunter did not address the issue of adverse impact using various predictor combinations.

Hattrup et al. (1997) examined validities and levels of adverse impact of a predictor composite while differentially weighting performance criteria. Specifically, they investigated the predictive validity of cognitive ability and work orientation for task and contextual performance in a Monte Carlo investigation of optimal weighting strategies for performance criteria. Hattrup et al. (1997) made the distinction between task and contextual performance and examined subgroup differences on the predictors given different weights applied to the criterion dimensions. As expected, cognitive ability was more related to task performance and work orientation was more related to contextual performance. Since work orientation is associated with less adverse impact, it was expected that weighting contextual performance higher on measures of job performance would weight the low impact predictor—work orientation—higher and yield less adverse impact in the predictor combination. However, the authors showed that a violation of the 80% rule was found at every selection ratio unless contextual performance was weighted five times higher than task performance, which would be highly unlikely in operational settings.

Although additional noncognitive constructs can add incremental validity beyond cognitive constructs (Ones et al., 1993; Schmitt et al., 1997), the empirical evidence suggests that combining cognitive predictors with these alternative predictors does not necessarily eliminate subgroup differences for a large range of selection ratios.

The consequences of adding alternative predictor constructs can vary widely and is based on many factors such as test validities, predictor intercorrelations, cutoff scores, specific stage of the selection process they are used, and selection ratios (Bobko et al., 1999; Huffcutt & Roth, 1998; Hunter & Hunter, 1984; Roth, Bobko, Switzer, & Dean, 2001; Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002; Sackett & Ellingson, 1997; Sackett & Roth, 1996; Schmitt et al., 1997). Therefore, one can conclude that it is difficult to enhance the validity of selection batteries while simultaneously reducing levels of adverse impact by simply using noncognitive predictors in combination with cognitively loaded predictors (Hough & Oswald, 2000; Hough et al., 2001; Ryan et al., 1998).

### *Changing the Method*

The second approach to reducing adverse impact recognizes that paper–and–pencil tests of knowledge, skill, ability and achievement are the most valid predictors of job performance but posits that racial subgroup differences on knowledge, skill, ability and achievement may partially arise from the mode or method of testing—specifically, paper–and–pencil multiple–choice tests—instead of the construct (Schmitt et al., 1996). Specifically, it is argued that the mode of testing influences test scores when it is not related to the criterion of interest (i.e., job or school performance; Ryan & Greguras, 1998). That is, racial test score differences are artifactual and result from contaminated predictor scores. Sources of artifactual variance include: the effects of test–taking skills that are not job related (Ryan & Greguras, 1998), the use of culture–specific language on the specified tests (Helms, 1992), and the use of test formats (e.g., multiple–choice) that discriminate among subgroups via processes of test

perceptions or reading demands that are not job related (Chan & Schmitt, 1997; Ryan & Greguras, 1998). Therefore, proponents of this theory of subgroup differences advocate the use of alternative testing formats such as video–based testing, job simulations, performance tests, and structured interviews that are not contaminated by job–irrelevant constructs such as reading load, susceptibility to testwiseness cues, or test perceptions.

A primary limitation of the extant literature is that it has often failed to draw a clear distinction between the method used to operationalize the construct and content (the construct measured) and the two have typically been confounded (Arthur et al., 2002; Chan, 1997; Chan & Schmitt, 1997; Schmitt et al., 1996). Schmitt et al. (1996) reviewed the literature on sex and race differences for some predictors of job performance such as spatial, verbal, and math abilities, job samples, interviews, and personality variables. They concluded that there was little research examining the effects of altering test methods on subgroup differences for job–relevant predictor constructs. In fact, they indicated that the results are difficult to interpret because the methodology used in most studies examining alternate predictors of job performance confounded the constructs and methods used to measure those constructs. That is, it is difficult to determine from the extant literature whether it is the alternate constructs that lower levels of adverse impact or alternate methods.

*Content* refers to the constructs and variables (e.g., conscientiousness, cognitive ability, finger dexterity, field dependence–independence, reaction time, and visual attention) that are being measured, and *method* refers to the techniques or procedures (e.g., graphology, paper–and–pencil tests, computer–administered tests, video–based tests, interviews, and simulations) that we use to accomplish the measurement of the

specified content (Arthur et al., 2003; Arthur et al., 2002; Campbell, 1990). From a methodological perspective, specified comparisons of different test formats should hold constructs constant and vary methods. For instance, if the levels of adverse impact for performance tests and paper–and–pencil multiple–choice tests are to be compared, then to obtain interpretable results, both test formats should be measuring the same construct. A comparison of paper–and–pencil measures of cognitive ability and work samples may not be meaningful if the constructs differ (Arthur et al., 2003; Arthur et al., 2002; Chan, 1997; Chan & Schmitt, 1997; Guion, 1998). Schmitt et al. (1996) found only one, albeit unpublished, study (Goldstein, Braverman, & Chung, 1992) that examined subgroup differences using different methods to measure the same constructs and called for more research using this methodology to examine subgroup differences on alternative testing methods. Since the Schmitt et al. review, we have found five studies (Arthur et al., 2002; Chan & Schmitt, 1997; Richman–Hirsch, Olsen–Buchanan, & Drasgow, 2000; Schmitt & Mills, 2001; Pulakos & Schmitt, 1996) that examined alternative testing methods holding the construct or content constant.

Pulakos and Schmitt (1996) represents an attempt at reducing subgroup differences by changing the method and holding the construct constant by measuring verbal ability and changing the method of measurement in a multi–predictor test battery. They argued that video–based testing more accurately reflects the cognitive demands encountered on the job. To test this theory, they compared scores on a traditional paper–and–pencil multiple–choice test and two job sample exercises. One job sample was a simulation of activities and tasks that would be encountered on a federal investigative job presented in a video–based format and participants were to

write an essay describing the activities they viewed in each simulation. The second job sample was a health fraud simulation in which participants were to write an essay documenting allegations of fraud by a doctor's patients. The health fraud simulation required participants to read written text describing allegations of fraud and procedures authorized by the accused doctor. African American–White subgroup differences were lower on the video–based administration (.45; Whites scored higher than African Americans) than on the written text medical fraud simulation (.91) and the multiple–choice test (1.03). Hispanic–White subgroup differences were also lower on the video–based administration (.37; Whites scored higher than Hispanics) than on the written text medical fraud simulation (.52), and the multiple–choice test (.78). The lower subgroup differences translated into lower levels of adverse impact against minorities for the video–based verbal ability test; however, the hiring rates based on the video administration alone still violated the 80% rule for African Americans at selection ratios of .60 and lower and for Hispanics at selection ratios of .40 and lower.

Chan and Schmitt (1997) compared subgroup differences on a video–based and written situational judgement test of work habits and interpersonal skills. They also examined participants' reports of the face validity of each test format and differences in reading comprehension as possible explanations for why subgroup differences were lower on video–based test formats. They found that performance on the video–based test was substantially higher than performance on the paper–and–pencil test, indicating that higher fidelity tests tend to yield increases in test performance for all participants. Furthermore, African American–White differences in test performance were also substantially smaller on the video–based test ($d$ = .21; Whites scored higher than

African Americans) compared to the paper–and–pencil test (*d* = .95). Additionally, when reading comprehension was controlled, there was a significant decease (i.e., 3% of the variance) in the race–test performance relationship. Finally, Whites reported higher levels of face validity for the paper–and–pencil test than African Americans (*d* = .80; Whites scored higher than African Americans), but that difference was significantly less for the video–based test (*d* = .11). Face validity accounted for a unique portion of the variance in subgroup differences (i.e., 3%) beyond that accounted for by reading comprehension.

Richman–Hirsch et al. (2000) compared three test methods that measured the same content for differences in examinees' perceptions (i.e., perceived fairness and face validity) and attitudes (i.e., enjoyment, shortness, satisfaction, and modernization). Participants reported more positive perceptions and attitudes towards the computerized, video–based multimedia version of a test that contained scenarios depicting workplace conflict as opposed to the written version of the same scenarios administered via a computer page–turner and paper–and–pencil formats. Participants perceived the computerized multimedia version to have more content and predictive validity and to be more job–related than the computer page–turner and paper–and–pencil formats. Their results suggest that newer, technologically advanced methods (e.g., computerized multi–media, video–based) of measuring valid constructs (i.e., cognitive ability) are viewed more favorably by examinees than written versions (e.g., paper–and–pencil, computerized page–turner tests).

Schmitt and Mills (2001) demonstrated that scores on a computerized job simulation yielded lower African American–White test score differences (*d* = .30;

Whites scored higher) than a paper–and–pencil test ($d$ = .61) measuring similar constructs. However, the computerized job simulation test had lower criterion–related validity (corrected $r$ = .36) than the paper–and–pencil test (corrected $r$ = .46). Furthermore, when the selection ratio was relatively low, the job simulation test still violated the 80% rule for adverse impact. Thus, although the job simulation method showed promising results as a viable alternative to traditional paper–and–pencil test formats, it did not completely eliminate the dilemma for personnel researchers and practitioners of finding a low adverse impact predictor with high criterion–related validity.

The evidence from the four studies reviewed above suggests that we can still use cognitive ability to select individuals into a wide variety of jobs, and also reduce racial subgroup differences when we use high fidelity testing formats such as multi–media presentations, performance tests, and job simulations. However, multi–media testing and job simulations have lower economic utility because they require more resources to develop (e.g., producing high–fidelity multi–media simulations and programming computer software), administer, and score (Chan & Schmitt, 1997). In addition to lower economic utility, applicant safety and a company's legal liability in the event of an accident also lowers the utility of performance tests (Hoffman & Thornton, 1997; Smither, Reilly, Millsap, Pearlman, & Stoffy, 1993; Weekley & Jones, 1997). Therefore, paper–and–pencil tests clearly have an advantage over performance tests and video–based or computerized testing, but the difficulty is constructing them to retain their advantages (e.g., mass administration, fewer required resources for administration

and scoring) and still minimize the alleged test bias typically associated with this methodology (Pulakos & Schmitt, 1996).

More recently, Arthur et al. (2002) compared multiple–choice and constructed response test formats that measured the same Fire Battalion Chief field management decision making abilities. Arthur et al. compared scores on a multiple–choice test to a constructed response test (write–in and mark–in format) in a within–subjects design and obtained a reduction of race–based subgroup differences on the constructed response format. Specifically, African American–White score differences and subsequent levels of adverse impact were lower for the constructed response ($d$ = .12; Whites scored higher on the test) than the multiple–choice test ($d$ = .70). In fact, the 80% rule for adverse impact was not violated with the constructed response format using any of the six different cutoff scores examined. In contrast, the 80% rule was violated in three of the six cutoff scores on the multiple–choice test. Finally, these results were also replicated using between–subjects data.

*Multiple–Choice vs. Constructed Response Tests*

Multiple–choice tests consist of an item stem and a set of alternatives from which to choose the correct answer to the stem. The advantages of multiple–choice tests include (a) the ability to administer the test to large numbers of people in one test administration, (b) they are relatively inexpensive to develop, administer, and score, (c) scoring is considered to be objective, and (d) a broad content domain can be represented on the test. In contrast, a constructed response format "is any question requiring the examinee to generate an answer rather than select from a small set of options" (p. 2, Bennett, 1993).

The item format of multiple–choice tests is fairly consistent, with only minor variations in stems and number of alternatives across tests. For example, stems may be written in question, sentence completion, or fill–in–the–blank form and the number of alternatives often ranges from two to five. In contrast, there is a broad range of constructed response item types that is defined by the degree of constraint on responses to the items. They can be defined along a continuum of constrained item types from sentence completion, to short answer (from one sentence to a short paragraph), to relatively unconstrained multi–page essays. Constructed response item formats are not commonly used in personnel selection but are seen more frequently in educational settings. However, given burgeoning objections to multiple–choice tests in high–stakes testing, constructed response formats may be a viable alternative. Indications of this orientation are seen in the emergent use of constructed response items on college admissions tests such as the SAT, GRE (Educational Testing Service, 2001a, 2001b), and GMAT (Graduate Management Admission Council, 2002) which now have writing assessment sections.

The problem however, is that by their very nature, constructed response tests are likely to produce less reliable scores and are also inherently more expensive to administer and score. For example, Bennett (1993) notes that "the costs of adding a single essay to a large–scale admissions testing program are about three to five times the cost of a 150–200–question multiple–choice test" (p. 11). Therefore, the challenge is to develop constructed response tests to retain the advantages of multiple–choice tests and ensure the construct equivalence of the two formats. Arthur et al. (2002) successfully developed a constructed response test with all four advantages of the

multiple–choice tests referred to above by using standard item–writing rules and a content–related validity approach. Arthur et al. argued that the two test formats measured the same construct since both tests were developed from the same content domain using a content–related validity approach. However, based on evidence that multiple–choice tests are often contaminated by other constructs (e.g., reading ability, test–taking skills) that are unrelated to job performance, it is not expected that the alternate form correlations will be unity.

As previously noted, Arthur et al. (2002) obtained a reduction in African American–White test score differences on a constructed response test compared to a multiple–choice test measuring identical content. However, Arthur et al. describe their data as suggestive due to the small sample sizes ($N = 27$ for the within–subjects data; $N = 51$ for the between–subjects data) and because the data were from an operational promotional exam in which the collection of non–operational research data or assignment to conditions was not possible. Thus, although Arthur et al. showed a reduction in subgroup differences and lower levels of subsequent adverse impact for their constructed response test, they provided no empirical data explaining these effects. However, as they discussed, there are several plausible reasons for the reduction in subgroup differences. Specifically, the use of constructed response tests may be associated with lower non–job–related reading load, lower susceptibility to testwiseness cues, and more favorable test perceptions than multiple–choice tests. Thus, in a replication of Arthur et al's findings, it was expected that:

H1: African American–White subgroup differences on the constructed response test will be smaller than African American–White subgroup differences on the multiple–choice test.

As previously noted, the use of constructed response tests may be associated with lower non–job–related reading load, lower susceptibility to test–taking skills, and more favorable test perceptions than multiple–choice tests. The next section reviews each of these factors, establishing the conceptual basis for the hypothesized subgroup differences and test–format effects.

### *Reading Ability*

In contrast to performance tests, the medium of information exchange for paper–and–pencil tests is via means of reading text and responding accordingly. To the extent that reading ability is not concomitant with job demands and/or performance, any variance associated with reading ability is considered to be error variance. That is, reading ability may be unrelated to job performance so that the measurement of reading ability in conjunction with job–relevant constructs (e.g., cognitive ability) reduces the construct validity of a selection or promotional test. Furthermore, to the extent that some evidence suggests the presence of racial subgroup differences in reading ability, then the reading load of the test medium (e.g., paper–and–pencil tests) may explain some of the variance in African American–White differences on test scores (Chan & Schmitt, 1997; Sacco, Scheu, Ryan, Schmitt, Schmidt, & Rogg, 2000). Indeed, the basic argument for the use of alternative test formats such as performance tests, job simulations, and tests with video–based presentation is that these tests have little or no

reading demands which translates into lower race–based subgroup differences (Chan & Schmitt, 1997; Schmitt & Mills, 2001).

Arthur et al. (2002) documented that the multiple–choice test used in their study contained 1,808 more words, 307 more sentences, and two more words per sentence than the constructed response test with comparable Flesch–Kincaid grade levels (10.99 for the multiple–choice and 10.50 for the constructed response). The observed differences in reading demands can clearly be attributed to the absence of alternatives on the constructed response test. Thus, the present study investigated differential reading loads as an explanatory variable for observed subgroup differences on an achievement test. Specifically, the reading load of a multiple–choice test is higher than a stem–equivalent constructed response test (Arthur et al., 2002). Consequently, it was expected that there will be a stronger relationship between reading ability and multiple–choice test performance than with constructed response test performance. Furthermore, based on the extant literature it was also expected that there would be subgroup differences in reading ability. Thus, when reading ability differences are controlled by using a constructed response test, race differences in test performance will be reduced or disappear.

H2a: There will be a significant, positive relationship between reading ability and test performance.

H2b: There will be a significant relationship between reading ability and test format. Specifically, the reading ability/multiple–choice relationship will be stronger than the reading ability/constructed response relationship.

H2c: There will be significant African American–White subgroup differences on reading ability.

H2d: Reading ability will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

*Test–Taking Skills*

It is widely recognized that performance on multiple–choice tests is influenced by the presence of testwiseness cues (Fagley, 1987; Millman, Bishop, & Ebel, 1965; Sarnacki, 1979; Traub, 1993). Millman et al. (1965) provided the first taxonomy to describe the construct of testwiseness which has served as the basis for the subsequent measurement and research of testwiseness. Testwiseness is defined as ". . . a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score. Testwiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures." (p. 707, Millman et al., 1965).

Both multiple–choice and constructed response tests are susceptible to test–taking skills, but the type of skills used on each test are very different. For example, essays require a test taker to generate a logical, clear, and concise response to the question using strategies specific to this test format. In contrast, test–taking skills used on multiple–choice tests typically involve the identification of correct alternatives by using secondary cues that would make the alternative attractive in the absence of knowledge of the test content. Examples of secondary cues include direct and indirect

clues such as lengthy correct alternatives, specific, unrelated, or absurd distractors that may be quickly eliminated, and carefully bounded qualifiers in the correct alternative. The specific type of test–taking skills examined in the present study are those that are associated with multiple–choice tests.

Test–taking skills have been measured by asking test takers to describe the strategies they use on the tests. These statements are then coded by experimenters into categories and given a score to represent the total number of strategies used and the relative effectiveness of the strategies (Bruch, 1981). Another method is to have test takers endorse items on a checklist describing various test–taking strategies used on a test such as a subsample of those outlined by Millman et al. (1965). One of the more common approaches to measuring test–taking skills is to develop multiple–choice items measuring content that would be unknown to the average person, but that may be answered correctly when applying basic test–taking strategies (e.g., Bajtelsmit, 1975; Gibb, 1964; Woodley, 1973). Gibb's (1964) 70–item Experimental Test of Testwiseness was developed using this methodology and is generally considered to be the most comprehensive measure of test–taking skills with the best psychometric properties (Harmon, Morse, & Morse, 1996; Miller, Fagley, & Lane, 1988; Miller, Fuqua, & Fagley, 1990).

The literature clearly indicates that test–taking skills are highly trainable and meta–analyses comparing groups trained in test–taking skills to control groups show an improvement in test scores from .1 to .2 standard deviations for the trained groups (Bangert–Drowns, Kulik, & Kulik, 1983; Messick & Jungeblut, 1981; Powers, 1993; Samson, 1985). This capacity for examinees to develop test–taking strategies is the

basis for the success of coaching to increase standardized test scores (e.g., Princeton and Kaplan Reviews; Bargent–Drowns et al., 1983; Dolly & Williams, 1986; Millman et al., 1965; Samson, 1985). Test–taking skills have also been associated with testing–taking experience. It stands to reason that those individuals with more test–taking experience are likely to have developed more effective test–taking strategies, are more comfortable taking tests, have more positive reactions toward the testing experience, may feel less anxiety during testing situations, and will spend less time reading familiar, standardized instructions (Sarnacki, 1979).

There are very few studies examining race–based subgroup differences in test–taking skills. Comparisons of Americans to Indonesian and Chinese students show a distinct advantage in test–taking skills exhibited by the American students (Lo & Slakter, 1973; Millman & Setijadi, 1966; Wu & Slakter, 1978). In contrast, several studies examining subgroup differences between racial groups within the United States have found no African American–White differences in test–taking skills (Benson, Urman, & Hocevar, 1986; Diamond, Ayres, Fishman, & Green, 1976; Rogers & Yang, 1996) despite claims that this variable may explain some of the variance in race differences (Helms, 1992; Kalechstein, Kalechstein, & Doctor, 1981). Ellis and Ryan (1999) provided evidence that African Americans reported the use of more ineffective test–taking strategies than Whites, but there were no differences for effective test–taking strategies. Two possible explanations for hypothesized race differences in test–taking skills is that Whites are more experienced with taking multiple–choice tests or that Whites are more likely to enroll in professional test coaching programs (e.g., the Princeton Review). However, Ellis and Ryan (1999) reported higher rates of

participation in test coaching programs in a sample of African Americans compared to Whites which would preclude hypothesizing that differences in training may explain the African American–White differences in test scores.

Test–taking skills are typically associated with the use of multiple–choice tests and is related only to the testing format and *not* the construct being measured by the test. In addition, evidence shows that test–taking skills influences test performance and since these skills are not related to job demands, any variance associated with this extraneous variable is considered error variance. To the extent that there are subgroup differences in test–taking skills this variable may explain some of the variance in African American–White performance differences on multiple–choice tests.

Given the paucity of research examining race differences in test–taking skills, the present study explored this variable as a possible explanation for a reduction in African American–White differences on a constructed response test. The limited evidence indicates that African Americans and Whites do not differ in test–taking skills, but hypotheses were presented in the direction of the other variables in the study. Specifically, multiple–choice tests are more susceptible to testwiseness cues, so it was expected that there will be a stronger relationship between test–taking skills and multiple–choice test performance than with constructed response test performance. Next, it was expected that there will be African American–White subgroup differences on test–taking skills. Finally, it was expected that when differences in test–taking skills are controlled by using a constructed response test, race differences in test performance will be reduced.

H3a: There will be a significant, positive relationship between susceptibility to testwiseness and test performance.

H3b: There will be a significant relationship between susceptibility to testwiseness and test format. Specifically, the test–taking skills/multiple–choice relationship will be stronger than the test–taking skills/constructed response relationship.

H3c: There will be significant African American–White subgroup differences in test–taking skills.

H3d: Test–taking skills will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

### *Test Perceptions*

The influence of examinee perceptions on test performance has been increasingly studied in the last 15 years (Ryan & Ployhart, 2000) and driven by findings that various perceptions (e.g., face validity) are related to real outcomes such as test performance or decision to withdraw from the job application process. It is important to note that Smither et al. (1993) indicated that reactions to test characteristics such as face validity are very important to consider in personnel selection because they may influence (a) perceived organizational attractiveness or the likelihood that someone accepts a job offer, (b) the potential for legal action, and (c) test performance through decreased motivation and absence of qualified applicants. Some researchers have posited that racial differences in standardized test performance can be attributed to

subgroup differences in face validity, perceived fairness, perceived predictive validity, belief in tests, stereotype threat, and self–efficacy (Chan & Schmitt, 1997; Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Gilliland, 1994; Ryan, 2001; Steele, 1997; Steele & Aronson, 1995).

Ryan (2001) provides the most comprehensive review of the testing literature that attempts to explain race–based test score differences on cognitively loaded tests through processes of test perceptions. She presents a model in which test perceptions can indirectly influence test performance via cognitions, affect, or motivation during test administration. It is posited that negative perceptions introduce test–irrelevant cognitions, negative affect, and lowered motivation that detract from test–relevant behaviors or cognitions associated with increased test performance (Ryan, 2001).

The following section reviews specific test perceptions hypothesized to influence test performance. Where pertinent, the review also summarizes specified relationships among the test perceptions as suggested by the literature. However, for the purposes of this study, hypotheses are limited to only the relationship between the specified perception variables and the race/test performance relationship. Nevertheless, inter–relations between the perception variables were assessed and reported.

*Face Validity*

One of the most common test reactions measured in the testing literature is the perceived face validity of the test (Chan et al., 1997; Smither et al., 1993). Face validity refers to perceptions of the extent to which a test "looks like" it is measuring what it is supposed to be measuring. Some advantages of using face valid tests are that

they are associated with smaller subgroup differences, more positive reactions from examinees, and are also legally more defensible (Chan et al., 1997).

Ryan (2001) suggested that face validity could interfere with test performance via a withdrawal of effort, increased anger, and decreased self–efficacy. However, most studies examining this variable have measured the relationship of reactions to the tests and post–test applicant behaviors rather than the relationship between pretest face validity and test performance. One exception is Chan et al. (1997), who administered parallel forms of a cognitive ability test and measured perceptions of face validity and test–taking motivation between administrations. They found that perceptions of face validity and motivation were significantly related to performance on the second test.

Most research examining race differences in face validity of different selection tests has focused on the face validity of the construct, rather than the test format (Ryan & Greguras, 1998). Of the limited research on face validity perceptions of test format, studies show that applicants generally perceive performance tests, job simulations, and video–based tests to be more job–related and fair than traditional paper–and–pencil tests (Chan et al., 1997; Schmidt, Greenthal, Hunter, Berner, & Seaton, 1977; Schmitt & Mills, 2001). For example, Chan et al. (1997) showed that test–takers perceived video–based tests to be more face valid than paper–and–pencil tests.

Evidence that face validity mediates the relationship between race and test performance has been empirically documented. For instance, Schmitt and Mills (2001) report smaller race differences on a job simulation than a traditional paper–and–pencil test and concluded that the job simulation was perceived to be more face valid. Chan and Schmitt (1997) compared a paper–and–pencil and video–based format on the same

situational judgement test and found the largest subgroup differences in face validity on the paper–and pencil test (Whites reported higher levels of face validity than African Americans). These studies compared performance tests and job simulations to traditional paper–and–pencil tests, but test–takers may even distinguish different paper–and–pencil test formats in terms of specific test perceptions (Outtz, 1998).

There is enough evidence (e.g., Outtz, 1998) to suggest that multiple–choice and constructed response formats differ in terms of face validity perceptions. Thus, the present study investigated face validity as a possible explanation for observed subgroup differences on an achievement test. Specifically, participants should report lower levels of face validity on the multiple–choice compared to the constructed response test. Consequently, it was expected that there will be a stronger relationship between face validity and multiple–choice test performance than with constructed response test performance. Furthermore, it was expected that there will be African American–White subgroup differences in face validity. Thus, when differences in face validity perceptions are controlled by using a constructed response test, race differences in test performance will be reduced.

H4a: There will be a significant, positive relationship between face validity and test performance.

H4b: There will be a significant relationship between face validity and test format. Specifically, the face validity/multiple–choice relationship will be stronger than the face validity/constructed response relationship.

H4c: There will be significant African American–White subgroup differences in face validity

H4d: Face validity will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

*Fairness*

Organizational justice theories are a useful framework to help explain applicant reactions to employment selection systems. Specifically, the perceived fairness of selection systems may be evaluated in terms of distributive and procedural justice (Gilliland, 1993; 1994). The majority of research examining the relationship between fairness and reactions to selection systems has typically examined post–test reactions and their influence on post–test behaviors such as recommendation intentions (Gilliland, 1994; Smither et al., 1993), acceptance intention (Macan, Avedon, Paese, & Smith, 1994) and withdrawal from the selection process (Schmit & Ryan, 1997). In general, the results of this research suggests that job applicants who perceive an overall selection system to be unfair will be less likely to recommend the employer to others or to accept a job offer, and they are also more likely to withdraw from the selection process.

Fairness may be related to performance on a specific test within a selection system. For example, tests perceived as unfair may result in a withdrawal of effort by the applicant. Some recent research in the area of applicant perceptions has examined the relationship between applicant pretest fairness perceptions and performance on cognitively loaded tests with somewhat mixed results. In terms of practical significance, fairness perceptions appear to account for only a small amount of variance

in test performance. For example, Chan, Schmitt, Sacco, and Deshon (1998) reported a small relationship between pretest fairness perceptions and average performance on three paper–and–pencil cognitive ability tests (mean $r$ = .23). Macon et al. (1994) obtained a similar relationship ($r$ = .21) between perceptions of fairness and an ability test battery. Finally, Ryan, Sacco, McFarland, and Kriska (2000) obtained an even smaller correlation between fairness and test performance ($r$ = .10).

The literature suggests that African American–White subgroup differences on fairness perceptions may also be quite small. For example, Chan et al. (1998) obtained an African American–White standardized difference of $d$ = .24 (Whites perceived the tests as being more fair) on three paper–and–pencil cognitive ability tests and Ryan et al. (2000) reported an African American–White difference of $d$ = .18 on an ability test. Nevertheless, fairness perceptions were examined in the present study as a possible explanation for observed subgroup differences on an achievement test. Specifically, it was anticipated that participants will report lower perceptions of fairness on the multiple–choice compared to the constructed response test. Consequently, it was expected that there will be a stronger relationship between fairness perceptions and multiple–choice test performance than with constructed response test performance. Next, it was expected that there will be African American–White subgroup differences in fairness perceptions. Finally, when differences in fairness perceptions are controlled by using a constructed response test, race differences in test performance will be reduced.

H5a: There will be a significant, positive relationship between fairness perceptions and test performance.

H5b: There will be a significant relationship between fairness perceptions and test format. Specifically, the fairness perceptions/multiple–choice relationship will be stronger than the fairness perceptions/constructed response relationship.

H5c: There will be significant African American–White subgroup differences in fairness perceptions.

H5d: Fairness perceptions will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

*Perceived Predictive Validity*

Perceived predictive validity refers to a test–taker's perception of how well test performance predicts future job performance. Smither et al. (1993) found evidence that perceptions of predictive validity of a selection procedure affects outcomes such as the willingness to recommend a company to others one month after completion of an employment examination. Other researchers contend that perceptions of predictive validity may influence test performance (Cascio, 1987; Chan et al. 1997) and Chan et al. (1997, 1998) hypothesized that this relationship is mediated by motivation to perform on the test. That is, low levels of perceived predictive validity may be associated with lower levels of test–taking motivation which directly influences test performance.

Several researchers have failed to find substantial race differences in the perceived predictive validity of cognitively loaded tests (Chan et al. 1998; Horvath,

Ryan, & Steirwalt, 2000; Smither et al. 1993). In one exception, Chan (1997) found an African American–White standardized difference of $d$ = .36 with Whites reporting higher predictive validity perceptions than African Americans on a paper–and–pencil cognitive ability test. However, Chan (1997) reported that the difference in predictive validity perceptions accounted for only a small reduction in performance differences (change in $d$ = .03).

No research studies have compared the perceived predictive validity associated with two different paper–and–pencil test formats. Therefore, the present study examined the role of perceived predictive validity as a partial explanation for observed subgroup differences on an achievement test. Specifically, it was anticipated that participants will report lower levels of perceived predictive validity on the multiple–choice compared to the constructed response test. Consequently, it was expected that there will be a stronger relationship between perceived predictive validity and the multiple–choice test performance than with constructed response test performance. Next, it was expected that there will be African American–White subgroup differences in perceived predictive validity. Finally, when differences in perceived predictive validity are controlled by using a constructed response test, racial subgroup differences in test performance will be reduced.

H6a: There will be a significant, positive relationship between perceived predictive validity and test performance.

H6b: There will be a significant relationship between perceived predictive validity and test format. Specifically, the perceived predictive

validity/multiple–choice relationship will be stronger than the perceived predictive validity/constructed response relationship.

H6c: There will be significant African American–White subgroup differences in perceived predictive validity.

H6d: Perceived predictive validity will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

*Belief in Tests*

Belief in tests is relevant to personnel testing because it refers to the strength of the belief that tests are valid indicators of job performance (Arvey, Strickland, Drauden, & Martin, 1990; Chan et al., 1998; Lounsbury, Bobrow, & Jensen, 1989; Ryan, 2001; Schmit & Ryan, 1992). This variable differs from the other perceptions reviewed in that it is a belief about all personnel selection tests and is not in reference to a specific test. Chan et al. (1998) showed that belief in tests has an indirect relationship with test performance through test characteristics such as face validity, perceived predictive validity, and fairness. However, Schmit and Ryan (1998) found no relationship between belief in tests and test performance.

The present study investigated belief in tests as a possible explanation for observed subgroup differences on an achievement test. Participants in the present study were informed that they would take a standardized achievement test similar to the ACT (American College Testing, 2002) exam used to predict college GPA. It was expected that this instruction would prime the widely held stereotype that standardized

achievement test scores are lower for African Americans than Whites. This stereotype should lower the belief that tests are valid indicators of academic performance for African Americans and this stereotype will be more pronounced on the multiple–choice test since this is the most widely used format for standardized testing. It was expected that there will be a stronger relationship between belief in tests and multiple–choice test performance than with constructed response test performance. It was also expected that there will be African American–White subgroup differences on belief in tests. Finally, when differences in belief in tests are controlled by using a constructed response test, race differences in test performance will be reduced.

H7a: There will be a significant, positive relationship between belief in tests and test performance.

H7b: There will be a significant relationship between belief in tests and test format. Specifically, the belief in tests/multiple–choice relationship will be stronger than the belief in tests/constructed response relationship.

H7c: There will be significant African American–White subgroup differences on belief in tests.

H7d: Belief in tests will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

*Stereotype Threat*

Stereotype threat is a theory for explaining the behavior and outcomes of persons in situations in which an identifying, specific sociocultural stereotype is highly

salient. Stereotype threat is described by Steele and Aronson (1995) as a person's "being at risk of confirming, as self–characteristic, a negative stereotype about one's group" (p. 797). Stereotype threat is a self–evaluative threat that may operate for members of any social–identity group defined by a widely held sociocultural stereotype to which its' members readily identify. The threat originates from one's personal identification with a negative stereotype and subsequent fear of fulfilling the stereotype or appearing to others as an example of the stereotype. The resulting fear or self–evaluative threat can substantially interfere with behavior. Steele and Aronson (1995) originally proposed this theory to explain lower performance of African Americans on cognitively loaded tests. Specifically, they indicated that African Americans are threatened by the risk of confirming the negative stereotype that African Americans score lower than Whites on standardized tests of knowledge, skill, ability, and achievement. This threatening condition negatively affects anxiety levels and motivation on a test which directly decreases test performance. Steele and Aronson (1995) further explain that as a mechanism to preserve their self–worth and self–esteem, African American students may actually devalue academic achievement thereby decreasing their investment in education and further exacerbating African American–White test score differences.

The theory as applied to testing simply refers to increased anxiety and decreased motivation in minorities in a high–stakes testing situation due to widely held stereotypes about the minority group to which the test–taker belongs. That is, the test–taker may not internalize or accept the veracity of the stereotype, but the knowledge that it exists for his/her specific race is enough to stimulate the threatening

condition. For example, Steele and Aronson (1998) found that presenting a "difficult test as diagnostic of ability" produced enough threat due to racial stereotypes in academically successful African Americans at Stanford University to disrupt performance on a cognitive ability test.

Evidence suggests the theory not only explains a *decrease* in test scores for a specified subgroup as a result of negative race–based stereotypes, but may also explain an *increase* in test scores for a specified subgroup as a result of positive race–based stereotypes. For example, in a study examining the susceptibility of children to stereotypes, Ambady, Shih, Kim, and Pittinsky (2001) found that childrens' test performance may be influenced by priming sociocultural stereotypes. For instance, lower elementary school (i.e., K–2$^{nd}$ grade) and middle school (grades 6–8) Asian–American girls performed significantly better than a control group on a math test when their ethnic identity was primed and significantly worse than the control group when their sex identity was primed. Presumably, when ethnic identity was primed the girls performed better on the math tests as a result of identifying with the stereotype that Asians are superior to other ethnic/racial minority groups on math. In contrast, when sex identity was primed the Asian–American girls scored worse than boys because they identified with the stereotype that girls are inferior to boys on math. Ambady et al. (2001) provided additional evidence that test performance may be susceptible to the influence of primed stereotypes in a positive direction. For example, they found that lower elementary school and middle school Asian–American boys performed significantly better on a math test than a control group with no identity priming when both the ethnic and sex stereotypes were activated. The implications are that children

as young as 5 years old are susceptible to prevalent sociocultural stereotypes such as males are superior in math compared to females and Asian Americans are superior in math compared to other ethnic groups.

The present study examined the role of stereotype threat on test performance to partially explain observed subgroup differences on an achievement test. Participants in the present study were informed that they would be taking a standardized achievement test similar to the ACT used to predict college grade point average (GPA). It was expected that this instruction set would prime the widely held stereotype that standardized test scores are lower for African Americans than Whites. Specifically, it was believed that the constructed response test would present a less threatening condition than the multiple–choice test since constructed response tests are not currently associated with standardized testing. Consequently, it was hypothesized that there will be a stronger relationship between stereotype threat and multiple–choice test performance than with constructed response test performance. Next, it was expected that there will be African American–White subgroup differences in perceived stereotype threat. Finally, when differences in the threatening condition are controlled by using a constructed response test, race differences in test performance will be reduced.

H8a: There will be a significant, negative relationship between perceived stereotype threat and test performance.

H8b: There will be a significant relationship between stereotype threat and test format. Specifically, the stereotype threat/multiple–choice relationship

will be stronger than the stereotype threat/constructed response relationship.

H8c: There will be significant African American–White subgroup differences on perceived stereotype threat.

H8d: Stereotype threat will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

*Self–Efficacy*

Self–efficacy is a belief that one has the ability to perform a specific task successfully (Bandura, 1997). The relationship between self–efficacy and performance on many tasks is well established in the literature. For example, Sadri and Robertson (1993) reported a corrected correlation between self–efficacy and job performance of .40 and Ryan (2001) reported the results of research that shows a moderate relationship (.22 – .23) between self–efficacy and test performance. There is very little research on race differences on test–taking self–efficacy. One exception is Horvath et al. (2000) who obtained a .18 correlation between race and self–efficacy with African Americans reporting lower levels of self–efficacy than Whites on a cognitive ability test.

Sarason (1980) reviewed the literature on self–efficacy and anxiety in test–taking situations and reported a strong relationship between the two constructs. Sarason explains that very anxious individuals often worry about their future performance because they are attending to past failures and potential shortcomings.

Hence, test anxiety and self–efficacy may be related through feedback on past test performance.

Despite the dearth of research on subgroup differences on test–taking self efficacy, the large body of literature reporting a strong relationship between self–efficacy and many other criteria suggests that this variable is important for test performance. Therefore, the present study examined the role of self–efficacy on test performance in an attempt to partially explain observed subgroup differences on an achievement test. It was expected that there will be a stronger relationship between self–efficacy and multiple–choice test performance than with constructed response test performance. Next, it is expected that there will be African American–White subgroup differences in self–efficacy. Finally, when differences in self–efficacy are controlled using a constructed response test, race differences in test performance will be reduced.

H9a: There will be a significant, positive relationship between self–efficacy and test performance.

H9b: There will be a significant relationship between self–efficacy and test format. Specifically, the self–efficacy/multiple–choice relationship will be stronger than the self–efficacy/constructed response relationship.

H9c: There will be significant African American–White subgroup differences on self–efficacy.

H9d: Self–efficacy will partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test.

### *Criterion–Related Validity*

Finally, an attempt to resolve the organizational dilemma of developing valid predictors of job performance that display lower levels of adverse impact should also evaluate the criterion–related validity of said predictors. In fact, one limitation of the extant literature is that examinations of subgroup differences on alternative predictors of job performance have typically failed to provide any criterion–related validity evidence. Although a reliance on content–related validity evidence may not in and of itself be deficient, the additional demonstration of criterion–related validity would further bolster the efficacy and utility of the predictor. Consequently, the present study compared the criterion–related validity of the constructed response test to the multiple–choice test using self–reported GPA as the primary criterion measure. It was expected that the criterion–related validity for the constructed response test will be comparable to that of the multiple–choice test. Furthermore, it was hypothesized that the relationship between self–reported GPA and constructed response test performance for African Americans will be the same as that for Whites.

H10a: The multiple–choice/GPA relationship will be the same as the constructed response/GPA relationship.

H10b: The constructed response/GPA relationship for African Americans will be the same as the constructed response/GPA relationship for Whites.

### *Summary of Research Objectives*

Using a college student sample, the present study attempted to replicate the results of Arthur et al. (2002) with a larger sample size and also empirically examine factors that may explain the reduction in race–based subgroup differences observed on

the constructed response compared to the multiple–choice test format. First, it was expected that race–based subgroup differences will be lower on the constructed response than the multiple–choice test. The present study also measured several variables—reading ability, test–taking skills, face validity, fairness, perceived predictive validity, belief in tests, stereotype threat, and self–efficacy—that may explain why subgroup differences are smaller on the constructed response test. In general, it was hypothesized that there are racial subgroup differences in reading ability, test–taking skills, and test perceptions. Furthermore, it was expected that these variables are related to test performance. In contrast to constructed response tests, multiple–choice tests are more susceptible to the operation of these variables. Therefore, observed subgroup differences in test performance may be fully or partially explained by the observed subgroup differences on these extraneous variables. Consequently, eliminating or controlling for the operation of these variables should reduce subgroup differences on multiple–choice tests to a level similar to that observed for the constructed response test. An illustration of the conceptual framework for the present study is presented in Figure 1. Much of this framework was derived from a conceptual model described by Ryan (2001).

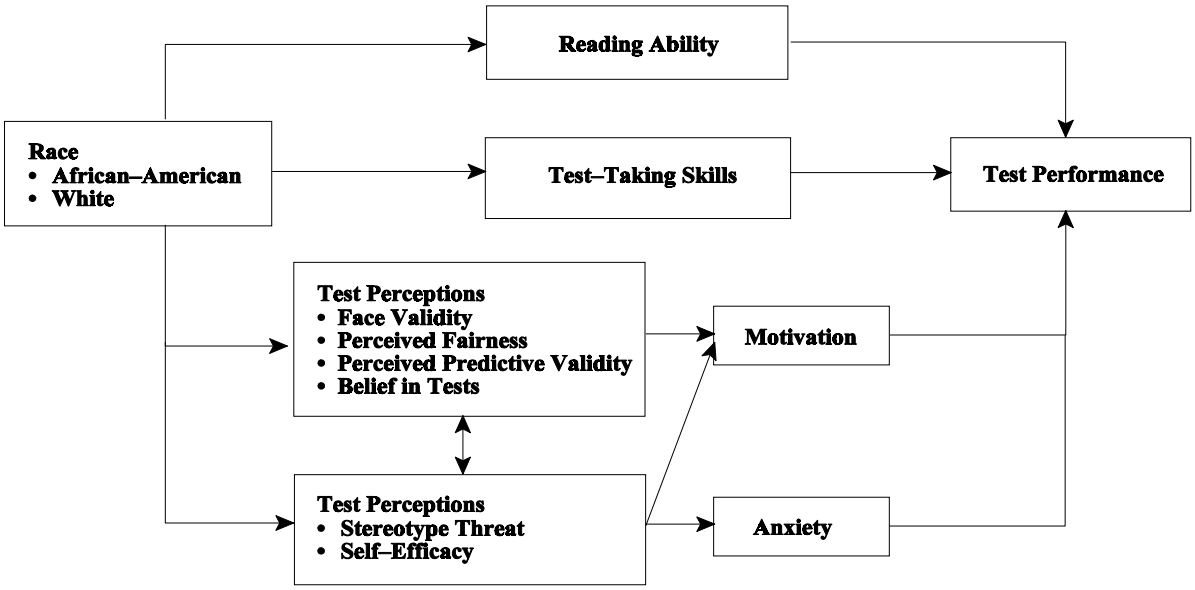


*Figure 1*. Model of the relationship between reading ability, test–taking skills, and test perceptions and multiple–choice and constructed response test performance.

# METHOD

## *Participants*

Participants were recruited from psychology department subject pools, upper–level undergraduate courses, and the general campus using posters, fliers, and newspaper advertisements from Texas A&M University ($n$ = 271), Prairie View A&M University ($n$ = 133), and the University of Houston ($n$ = 60). The mean age for the sample was 20.1 ($SD$ = 2.1). Sample sizes for each condition by race and sex are presented in Table 1.

Table 1
*Sample Sizes for Each Condition by Race and Sex*

| | Multiple–Choice | | Constructed Response | | |
|---|---|---|---|---|---|
| Sex | African American | White | African American | White | **Total** |
| Males | 34 | 40 | 16 | 54 | **144** |
| Females | 81 | 67 | 73 | 99 | **320** |
| **Total** | **115** | **107** | **89** | **153** | **464** |

Participants recruited from the subject pool received research credit in introductory psychology; non–subject pool participants received either extra course credit or $15 for their participation. There was no significant difference ($t$ [462] = 1.12, $ns$) in reported level of effort exhibited on the tests as a function of type of compensation (i.e., participants that were paid vs. participants that received extra course credit or research credit in introductory psychology). There were also no differences between the two types of compensation and participant scores on any of the study variables with the exception of age ($t$ [462] = 10.0, $p < .001$; participants' receiving

payment mean age = 20.7, *SD* = 2.32; participants' receiving extra course credit or research credit mean age = 19.1, *SD* = 1.20). However, this difference was expected as the majority of participants in the latter category received research credit because they were in introductory psychology classes (*n* = 180) which consist mostly of underclassmen (e.g., Freshmen and Sophomores) who are typically younger students. To motivate research participants to exert effort on the multiple–choice and constructed response tests, participants with the 20 highest scores were awarded $30.

Power analyses (Cohen, 1988) were performed for the hypotheses with the most conservative test (i.e., multiple regression with a covariate and two independent variables). Specifically, regressing the dependent variable test scores on to face validity (a covariate with the weakest relationship with test performance), race, and test format yielded multiple $R^2$ = .30 and power of 1.00. Therefore, the power of the present study was adequate for all subsequent analyses that involved larger effect sizes or fewer independent variables.

## *Measures*

### *Cognitive Ability*

The Raven's Advanced Progressive Matrices Short Form (APM; Arthur & Day, 1994; Raven, Raven, & Court, 1994) is a measure of general cognitive ability which consists of 12 matrix or design problems arranged in an ascending order of difficulty and is scored by summing the number of problems answered correctly. The short form of the APM (Arthur & Day, 1994; Arthur, Tubré, Paul, & Sanchez–Ku, 1999) was used in the present study to control for any preexisting differences between groups on ability level. The APM short form demonstrates psychometric properties similar to that of the

long form with a reduced administration time of 15 minutes. The odd/even split–half reliability with a Spearman–Brown correction for the APM scores was .67.

*Reading Ability*

Reading ability was measured in the present study using the comprehension subtest of the Nelson–Denny Reading Test (Brown, Bennet, & Hanna, 1993, Form G). The comprehension subtest in Form G consists of seven passages of text and 38, 5–alternative multiple–choice items. Administration time is 20 minutes and the technical report provides test–retest reliabilities of .75 to .82 for scores on the comprehension subtest. The test has a high school to college reading level and is used extensively in educational settings. The data obtained from the pilot study (described below in a later section) was used to reduce the number of items on the Nelson–Denny Reading Test to 20 items. Consequently, the reduction of items resulted in a reduced administration time of 10 minutes. In addition, because reading ability was one of the primary variables of interest in this study, half of the items on the Nelson–Denny Reading Test were administered in a constructed response format. This was done to balance the response formats and prevent artificial inflation of the reading ability/multiple–choice test performance relationship due to common method variance. For the reading ability test, the internal consistency was .65 for the multiple–choice scores, .67 for the constructed response scores, and .79 for the full 20–item test scores.

*Test–Taking Skills*

Gibb's (1964) Experimental Test of Testwiseness is generally considered to be the most comprehensive measure of test–taking skills with the best psychometric properties (Harmon et al., 1996; Miller et al., 1988; Miller et al., 1990). The

Experimental Test of Testwiseness measures a content domain largely unknown to the general public (i.e., obscure history facts and interpretations). The test consists of 70 multiple–choice items that, in the absence of knowledge concerning the history content, can only be answered using secondary, format–related cues, intentionally written into the items so that highly skilled test–takers will be directed to the correct answer through knowledge and application of one of 7 types of testwiseness cues. Administration time is 25–30 minutes. Gibb (1964) reported a KR–20 reliability of .72 for scores on the Experimental Test of Testwiseness. Data from the pilot study (described below in a later section) was used to reduce the number of items on the Experimental Test of Testwiseness to 20 items for an administration time of 10 minutes. The shortened version of the measure is presented in Appendix A. The internal consistency for the test scores was .57.

*Standardized Test Coaching*. The present study used one item that asked participants to indicate if they have ever taken a coaching class for standardized tests (e.g., the Princeton Review) to examine its influence on test performance. The item is in the post–test questionnaire presented in Appendix B.

*Experience with Tests*. Prior experience with taking multiple–choice (1 item) and constructed response tests (1 item) was measured in the present study to determine it's influence on test performance and as a check on the convergent validity of the measure of test–taking skills. Furthermore, researchers have strongly suggested measuring experience with test–taking when measuring reactions to test formats (Ryan & Greguras, 1998). The experience with tests items are in the post–test questionnaire presented in Appendix B.

*Face Validity*

Face validity was measured using four items adopted from Smither et al. (1993). Examples of items are "I cannot see any relationship between this test and what is required in college courses" and "It would be obvious to anyone that this test is related to college performance." Ratings were made on a 5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the face validity ratings was .75 in the multiple–choice test condition, .82 for the constructed response test condition, and .79 across both conditions. The face validity items are in the Test Attitudes and Perceptions Survey presented in Appendix C.

*Fairness*

Perceived fairness was measured using five items adopted from Smither et al. (1993). Examples of items are "The test results will accurately reflect how well I perform on this test" and "I deserve the test results that I will receive on this test." Ratings were made on a 5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the fairness ratings was .65 in the multiple–choice test condition, .63 in the constructed response test condition, and .64 across both conditions. The fairness items are in the Test Attitudes and Perceptions Survey presented in Appendix C.

*Perceived Predictive Validity*

Perceived predictive validity was measured using four items adopted from Smither et al. (1993). Examples of items are "Failing to pass this test clearly indicates that you can't pass many college courses and "I am confident that the test can predict how well an applicant will perform in college courses." Ratings were made on a

5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the perceived predictive validity ratings was .78 in the multiple–choice test condition, .86 in the constructed response test condition, and .82 across both conditions. The perceived predictive validity items are in the Test Attitude and Perceptions Survey presented in Appendix C.

*Belief in Tests*

Belief in tests was measured using four items adopted from the Test Attitude Survey (TAS; Arvey et al., 1990) which is a general measure of employment test–taking attitudes and opinions that may be applied to achievement tests (i.e., ACT) for selection into colleges. The original TAS is a 45–item measure of dispositional test anxiety which measures 9 attitudes toward tests: (1) motivation to perform well on the test, (2) lack of concentration on the test, (3) belief in tests in general, (4) comparative anxiety, (5) test ease, (6) external attribution, (7) general need achievement, (8) future effects, and (9) preparation. The TAS was written to assess reactions to a test after the administration of the test. Therefore, items from Scale 3 were reworded to measure a belief in tests *before* the multiple–choice or constructed response test was administered to obtain a measure of pretest perceptions. Examples of items are "Tests are a good way of selecting people for college" and "I don't believe that tests are valid." Ratings were made on a 5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the belief in tests ratings was .69 in the multiple–choice test condition, .70 in the constructed response test condition, and .69 across both conditions. The belief in tests items are in the Test Attitudes and Perceptions Survey presented in Appendix C.

*Stereotype Threat*

Stereotype threat was measured using six items adopted from Steele and Aronson (1995) that ask participants about their knowledge of the stereotype that standardized tests of ability are biased against African Americans and the degree to which they identify themselves as "good test–takers". An example of an item is "I feel self–conscious when taking tests." Ratings were made on a 5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the stereotype threat ratings was .43 in the multiple–choice test condition, .50 in the constructed response test condition, and .47 across both conditions. The stereotype threat items are in the post–test questionnaire presented in Appendix B.

*Self–Efficacy*

Self–efficacy was measured using three items adopted from Arthur, Bell, and Edwards (2003). The items were developed following principles and guidelines recommended by Bandura (1997) for the development of self–efficacy scales. An example item is "I believe I will perform well on this test." Ratings were made on a 5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the self–efficacy ratings was .75 in the multiple–choice test condition, .82 in the constructed response test condition, and .83 across both conditions. The self–efficacy items are in the Test Attitudes and Perceptions Survey presented in Appendix C.

*Test–Taking Motivation*

Test–taking motivation was measured using eight items adopted from the TAS (Arvey et al., 1990). Items from Scale 1 were reworded to measure test–taking

motivation *before* the multiple–choice or constructed response test was administered. Examples of items are "Doing well on this test is important to me" and "I will try my best on this test". Ratings were made on a 5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the test–taking motivation ratings was .91 in the multiple–choice test condition, .90 in the constructed response test condition, and .91 across both conditions. The test–taking motivation items are in the Test Attitudes and Perceptions Survey presented in Appendix C.

*Level of Effort in the Study*

As a check on the participants' level of effort in the study, one item was administered at the end of the protocol to measure participants' motivation to perform well on all tests in the protocol using a 5–point scale (1 = no effort, 5 = a lot of effort). The item is in the post–test questionnaire presented in Appendix B.

*Test–Taking Anxiety*

Test–taking anxiety was measured using 10 items adopted from the TAS (Arvey et al., 1990). Items from Scale 4 were reworded to measure test–taking anxiety *before* the multiple–choice or constructed response test is administered. Examples of items are "I am not good at taking tests" and "I usually get very anxious about taking tests". Ratings were made on a 5–point Likert scale (1 = strongly disagree; 5 = strongly agree). The internal consistency for the test–taking anxiety ratings was .84 in the multiple–choice test condition, .87 in the constructed response test condition, and .86 across both conditions. The test–taking anxiety items are in the Test Attitudes and Perceptions Survey presented in Appendix C.

*Dependent Variable*

*Multiple–Choice Test*. The multiple–choice test measured mathematics and science reasoning abilities using selected items from the ACT exam (American College Testing, 2002). The ACT consists of four subscales (English, mathematics, reading, and science reasoning). Only items from the mathematics and science reasoning subscales of the ACT were used in the present study. It is difficult to create a constructed response test using the English subscale from the ACT because there are multiple answers for each item and the reading subscale is of the same format as the Nelson–Denny Reading Test used in the present study as a measure of reading ability.

A subset of 20 items (10 math and 10 science reasoning) were selected from a previously released version of the ACT, preserving their ascending order of difficulty. Items on the math portion of the ACT are written in a five–alternative, multiple–choice format which were retained for the current multiple–choice test. Items on the science reasoning portion are written in a four–alternative, multiple–choice format which were retained for the multiple–choice test. The number of item types selected for both sections were similar in proportion to their frequency of occurrence on the original test. For the present study, an additional, non–keyed distractor was developed and added to the items in the science–reasoning section to provide consistency in the number of alternatives across the test sections.

Participants were given 25 minutes to complete both the math and science reasoning sections of the test. Pilot testing (described below in a later section) revealed that the 25–minute time limit was sufficient for the majority of test–takers to complete the test. The test was scored by summing the number of problems answered correctly.

The odd/even split–half reliability with a Spearman–Brown correction for the multiple–choice test scores was .83.

*Constructed Response Test*. Constructed response tests must be developed to limit the range of possible correct responses. That is, the range of responses should be constrained such that there will be only one correct response. This serves to minimize the subjectivity in scoring and increases reliability. The constructed response test items for the present study were designed to meet these criteria and were stem–equivalent to the multiple–choice test items to ensure maximum content overlap between the two tests. The stem–equivalent constructed response format used in the present experiment was similar to the write–in format used by Arthur et al. (2002). In fact, the constructed response test was simply the items on the multiple–choice test with the alternatives eliminated. Participants were required to provide the correct answer by writing it on a blank line under the stem.

Participants were given 25 minutes to complete both the math and science reasoning sections of the test. Pilot testing (described below in a later section) revealed that the 25–minute time limit was sufficient for the majority of test–takers to complete the test. The test was scored by summing the number of problems answered correctly. A scoring key was developed by the primary researcher that specified the correct responses for each item. Two scorers were trained on how to score the test, and applying this key to each test, scored items as correct or incorrect. Next, responses were entered into a text file for analysis. The two scorers independently scored 14 constructed response tests and the degree of overlap was 95.8%. The primary researcher and the two scorers then met to resolve the discrepancies and the scorers

independently scored an additional 10 tests in which the degree of overlap was 99.2%. The remainder of the constructed response tests were scored and entered by only one of the two scorers. Based on a sample of 60 tests, it took on average 55 seconds to score the constructed response test and 31 seconds to enter the scores in a text file. The odd/even split–half reliability with a Spearman–Brown correction for the constructed response test was .78.

*Practice Test*. A practice test was created by obtaining four multiple–choice and constructed response sample items from the math section and four multiple–choice and constructed response sample items from the science reasoning section of a previously released version of the ACT. These items provided participants with exposure to the content and format of their assigned test (multiple–choice or constructed response) so that accurate ratings of pretest perceptions could be obtained.

*Criteria*

Self–reported cumulative college GPA was used as the primary criterion (Cassady, 2001). Additional criterion measures such as overall high school GPA, and high school rank were also collected. The additional criterion measures were deemed necessary because a large proportion of the participants were college freshmen ($n$ = 132) who may not yet have a college GPA.

All criterion data were collected at the beginning (in the demographic questionnaire) and the end (post–test questionnaire) of the protocol. The purpose of this design was to test the "temporal stability" of using self–report to collect these data: college GPA ($r$ = .99), high school GPA ($r$ = .94), and high school rank ($r$ = .99).

In the pilot demographic and post–test questionnaires, participants were asked to report scores for college GPA, high school GPA, and high school rank. However, a large percentage (i.e., 21%) of the scores were missing. Feedback from some of the participants during pilot study data collection indicated that this was due to the format of the items which required participants to recall exact values. Consequently, the decision was made to provide a range of values on these items and have participants select a range instead of a specific, exact value. For example, college grade point average was provided on a 7–point scale (1 = 3.5 to 4.0; 2 = 3.0 to 3.4; 3 = 2.5 to 2.9; 4 = 2.0 to 2.4; 5 = 1.5 to 1.9; 6 = 1.0 to 1.4; 7 = Below 1.0). An additional option was provided for participants who did not yet have a GPA at their current university. The GPA range values were reversed scored such that higher scale values indicate a higher college GPA. The items for college GPA, high school GPA, and high school rank are in the demographic and post–test questionnaires presented in Appendices B and D.

*Self–Reported SAT/ACT Scores*

Self–reported scores on the SAT and/or ACT were also collected to assess the convergent validity of the multiple–choice and constructed response tests used in the present study. The SAT is the most widely used college entrance test in Texas so this was the test that was reported by most of the participants (301 reported SAT scores, 82 reported ACT scores, 41 reported both, and 40 did not report SAT or ACT scores). For those participants who reported both SAT and ACT scores, only the SAT scores were used in the analyses. Scores for the 82 participants who reported only ACT scores were converted to the SAT scale using the mean ($M$ = 1,115.56) and standard deviation ($SD$ = 177.65) obtained from the current sample.

SAT/ACT scores were collected at the beginning (in the demographic questionnaire) and the end (post–test questionnaire) of the protocol. The purpose of this design was to test the "temporal stability" of using self–report to collect these data ($r$ = .95). The items for the collection of self–reported SAT/ACT scores are in the demographic and post–test questionnaires presented in Appendices B and D.

## *Design and Procedure*

### *Pilot Study*

Given the nature of the research questions addressed in the present study, it was important that all data be collected in a single session. However, use of most measures in their original form required an estimated administration time of 2 ½ hours which was not practically feasible for a single data collection session. As an alternative, multiple data collection sessions were also not feasible due to the possibility of high attrition rates. Thus, a pilot study was conducted to shorten the measures, refine those developed for the study, and develop a research protocol that could be administered in less than two hours. Therefore, the data from the pilot test were used to reduce the number of items on the Nelson–Denny Reading Test, Experimental Test of Testwiseness, and the Test Attitudes and Perceptions Survey. In addition, time to completion was recorded for each measure and initially estimated time limits were reduced to more accurately reflect actual administration time.

To reduce the specified tests to their target lengths, the tests were administered to participants in the pilot testing phase of the study and a set of psychometric decision rules were applied. First, items with the highest item–total correlations (i.e., above .50) were retained and in the case of ties, the most difficult items were retained. To further

reduce test lengths, items with the strongest contribution to the test's coefficient alpha were retained until the specified test lengths were obtained. For the Nelson–Denny Reading Test and the Experimental Test of Testwiseness, the target test length was 20 items each. There was no *a priori* determined target test length for the Test Attitudes and Perceptions Survey, so the test length was reduced such that only the items with the best psychometric properties for each construct were retained, reducing the test length from 42 to 38 items. Finally, some items on the demographic and post–test questionnaires were revised or eliminated based on response rates, participants' verbal feedback, analysis of the data, and further evaluation of the items. Reducing the tests and time limits shortened the length of the protocol to 1 hour and 40 minutes.

*Present Study*

Table 2 presents the list of the measures, number of items and administration time both prior to and after the pilot study, and the order in which they were administered in the research protocol. The present study used a between–subjects design. Thus, participants were randomly assigned to either the multiple–choice test or constructed response test condition. Data collection was limited to a single session which lasted approximately 1 hour and 40 minutes.

Participants first read and signed the informed consent and provided demographic (i.e., age, race, sex, and college classification; see Appendix D) and criterion data. Next, participants received either the multiple–choice or constructed response practice test. Participants read the instruction set as it was read aloud by the proctor and then completed the practice test. Participants next completed the Test Attitudes and Perceptions Survey. The 20–item achievement test (multiple–choice or

constructed response) was then administered. Following completion of the multiple–choice or constructed response test, participants completed the APM, Nelson–Denny Reading Test, Experimental Test of Testwiseness, and the post–test questionnaire.

Table 2
*Study Measures*

| | Pre–Pilot | | Post–Pilot | | |
|---|---|---|---|---|---|
| Measure | # Items | Time (minutes) | # Items | Time (minutes) | Order of Administration |
| [A]Demographic Questionnaire | 5 | 3 | 8 | 3 | 1 |
| [B]Practice Test | 8 | 10 | 8 | 8 | 2 |
| [B]Test Attitudes and Perceptions Survey | 42 | 20 | 38 | 8 | 3 |
| [B]Multiple–Choice or Constructed Response Test | 50 | 30 | 20 | 25 | 4 |
| [C]Raven's Short Form | 12 | 15 | 12 | 15 | 5 |
| [B]Nelson–Denny Reading Test | 38 | 25 | 20 | 10 | 6 |
| [B]Experimental Test of Testwiseness | 70 | 40 | 20 | 10 | 7 |
| [A]Post–test Questionnaire | 11 | 5 | 9 | 5 | 8 |
| **TOTAL** | **236** | **148** | **135** | **84** | |

Note: [A]Measure was created for the present study; [B]Standardized test that was altered for the purposes of the present study; [C]Standardized test.

One concern that arises when recruiting participants to participate in research where the outcomes have no real consequences is that their level of effort may be much lower. Therefore, the present study attempted to enhance participant motivation on the tests by rewarding $30 to the participants with the highest 20 scores on the tests. Furthermore, an item was administered at the end of the protocol that assessed the level of effort put forth on the tests as a check of motivation. Participants were urged to do their very best and to complete each test and survey in its entirety.

*Replacement of Missing Data*

To equate sample sizes for all of the analyses and thus guard against a loss of statistical power and bias in the parameter estimates, missing data were replaced using a multiple regression imputation approach. Listwise deletion was the least desirable approach to deal with missing data because it sacrifices a large amount of data. Therefore, regression imputation was used because it is appropriate for replacing missing data when the data are missing both randomly and nonrandomly and when 10% or less of the data are missing (Roth, 1994). In addition, it is superior to other missing data techniques such as listwise deletion, pairwise deletion, and mean substitution. For each variable containing missing data, variables with the largest zero–order correlation with the missing variable were selected for the regression equation to maximize $R^2$ using scores from only the participants with complete data. Next, variables were retained in the regression that yielded a significant, unique contribution to the prediction of the missing variable score as defined by a significant $\Delta R^2$ when introduced into the equation. Finally, the participant's data for the missing variable(s) was replaced with the predicted score by applying the regression weights to the participant's known

scores. It is important to note that predictor variables were not used to impute criterion variables and vice versa as this would artificially inflate the relationships under investigation. In addition, missing scores on variables related to perceptions of and reactions to the specific tests were imputed using regression weights derived only from the appropriate test condition (i.e., multiple–choice and constructed response).

Among the predictor variables, one participant (< 1%) was missing data for face validity and perceived predictive validity, two participants (< 1%) were missing data for perceived fairness, three participants (< 1%) were missing data for test–taking motivation, and 38 participants (8%) were missing data for stereotype threat. Part of the sample (i.e., $n$ = 108) were administered the protocol in a classroom setting and due to time constraints the measure of test–taking strategies was omitted from the protocol. Therefore, 108 participants (23%) were missing data on this variable and scores were not imputed. As such, all analyses using test–taking skills were restricted to a sample of 356. All other analyses were based on an $N$ of 464.

Among the criterion variables, 27 participants (6%) were missing data for college GPA, one participant (< 1%) was missing data for high school GPA, and four participants (< 1%) were missing data for high school rank. Finally, 40 participants (9%) were missing data for SAT/ACT scores.

Although they did not exceed the general rule–of–thumb, examination of the reading ability data indicated that they were negatively skewed (skewness = -1.43). Thus, the decision was made to transform the data using a cubed transformation before performing parametric tests. The means and standard deviations of raw scores for

reading ability are reported to allow direct interpretation of results. However, all parametric tests were based on the transformed reading ability data.

# RESULTS

Descriptive statistics and correlations for all study variables are provided in Table 3 and for only the hypothesis–related variables in Table 4. Variables related to test perceptions (i.e., face validity, fairness, perceived predictive validity, belief in tests, stereotype threat, self–efficacy, motivation, and anxiety) were measured following the practice test and prior to the multiple–choice or constructed response test. Therefore, ratings were made in reaction to having been exposed to one of the two testing conditions. As such, the intercorrelations among these variables are presented separately for both test formats.

## *Convergent Validity of the Multiple–Choice and Constructed Response Tests and SAT Scores*

The multiple–choice and the constructed response test were developed using selected items from the ACT which is often used in selection for college and university admissions. Therefore, scores on the multiple–choice and constructed response tests should be significantly correlated with self–reported scores for the SAT, another test used in selection for colleges and universities. As seen in Table 4, SAT scores were significantly related to scores on both the multiple–choice ($r = .62$, $p < .001$, 95% CI = .53 to .69) and the constructed response ($r = .65$, $p < .001$, 95% CI = .57 to .72) test, demonstrating convergent validity for these two tests.

## *Standardized Test Coaching*

The present study asked participants to indicate if they had ever taken a coaching class for standardized tests (e.g., the Princeton Review) to examine its influence on test performance. The results indicate that there was a significant positive

relationship between number of standardized test coaching courses taken and test performance ($r$ = .21, $p$ < .001, 95% CI = .12 to .30). In addition, the multiple–choice/coaching and constructed response/coaching relationships were identical ($r$ = .21, $p$ < .001, 95% CI = .09 to .33). Whites ($M$ = 5.48, $SD$ = .89) reported taking more coaching courses than African Americans ($M$ = 4.99, $SD$ = 1.14; $t$ [461] = 4.97, $p$ < .001; $d$ = .48).

*Experience with Tests*

Prior experience with taking both multiple–choice and constructed response tests was measured in the present study to determine its relationship with test performance and as a check on the convergent validity of the measure of test–taking skills. First, the relationship between prior experience with taking multiple–choice tests and performance on the multiple–choice test in the present study was not significant ($r$ = -.07, *ns*, 95% CI = -.20 to .06) Likewise, the relationship between prior experience with taking constructed response tests and performance on the constructed response test ($r$ = .12, , *ns*, 95% CI = .01 to .24) was also not significant. Therefore, experience with taking multiple–choice and constructed response tests was not related to test performance. Although there were no subgroup differences in experience with taking multiple–choice tests ($t$ [461] = .26, *ns*), Whites ($M$ = 3.21, $SD$ = 1.12) reported more experience with taking constructed response tests than African Americans ($M$ = 2.82, $SD$ = 1.19; $t$ [459] = 3.64, p < .001; $d$ = .34). There was no relationship between

Table 3

*Descriptive Statistics and Correlations for All Study Variables*

| Variable | *M* | *SD* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. MC[A] Test | 13.40 | 4.18 | (83) | | | | | | | |
| 2. CR[B] Test | 11.11 | 4.24 | — | (78) | | | | | | |
| 3. Race[C] | — | — | 56 | 43 | — | | | | | |
| 4. Cognitive Ability | 7.44 | 2.40 | 59 | 52 | 41 | (67) | | | | |
| 5. Reading Ability | 17.04 | 2.99 | 55 | 50 | 42 | 35 | (80) | | | |
| 6. Test–taking Skills | 8.60 | 3.10 | 33 | 36 | 21 | 27 | 30 | (57) | | |
| 7. Standardized Test Coaching | 5.26 | 1.04 | 21 | 21 | 23 | 15 | 21 | 08 | — | |
| 8. Experience with MC Tests | 3.81 | 1.84 | -07 | -04 | 01 | -05 | 00 | 03 | 22 | — |
| 9. Experience with CR Tests | 3.03 | 1.16 | 05 | 12 | 17 | 04 | 02 | -02 | 11 | 21 |
| 10. Face Validity | 2.79/2.70 | 0.70/0.75 | -04 | 00 | -10 | 01 | -06 | -04 | -04 | -05 |
| 11. Fairness | 3.12/3.12 | 0.57/0.52 | 30 | 17 | 12 | 22 | 15 | 11 | 02 | -03 |
| 12. Predictive Validity | 2.14/2.12 | 0.68/0.73 | 13 | 12 | 00 | 11 | 01 | 04 | -02 | -08 |
| 13. Belief in Tests | 2.60/2.66 | 0.73/0.68 | 15 | 19 | 00 | 13 | 09 | 01 | -04 | -03 |
| 14. Stereotype Threat | 2.77/2.74 | 0.48/0.52 | -27 | -41 | -08 | -25 | -19 | -11 | 02 | 05 |
| 15. Self–Efficacy | 3.28/3.28 | 0.78/0.79 | 22 | 35 | -12 | 23 | 13 | 16 | 00 | 02 |
| 16. Motivation | 3.92/3.93 | 0.71/0.67 | 02 | 09 | -19 | 06 | 01 | -05 | -02 | -03 |
| 17. Level of Effort | 2.69 | 0.95 | 28 | 16 | -06 | 17 | 10 | 16 | -07 | -04 |
| 18. Anxiety | 2.92/2.89 | 0.68/0.72 | -28 | -37 | 06 | -24 | -14 | -17 | -04 | -01 |

Table 3 Continued

| Variable | *M* | *SD* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| 19. College GPA | 5.44 | 1.11 | 13 | 20 | 05 | 10 | 09 | 07 | 04 | 03 |
| 20. High School GPA | 6.38 | 0.84 | 36 | 46 | 27 | 28 | 29 | 11 | 06 | -05 |
| 21. High School Rank | 4.58 | 1.86 | 30 | 27 | 21 | 17 | 21 | 11 | 02 | -03 |
| 22. SAT | 1096.90 | 167.07 | 62 | 65 | 51 | 55 | 50 | 36 | 21 | 02 |
| 23. Test Performance[D] | 12.21 | 4.36 | — | — | 43 | 50 | 48 | 32 | 21 | -06 |

Table 3 Continued

| Variable | *M* | *SD* | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. MC[A] Test | 13.40 | 4.18 | | | | | | | | |
| 2. CR[B] Test | 11.11 | 4.24 | | | | | | | | |
| 3. Race[C] | — | — | | | | | | | | |
| 4. Cognitive Ability | 7.44 | 2.40 | | | | | | | | |
| 5. Reading Ability | 17.04 | 2.99 | | | | | | | | |
| 6. Test–taking Skills | 8.60 | 3.10 | | | | | | | | |
| 7. Standardized Test Coaching | 5.26 | 1.04 | | | | | | | | |
| 8. Experience with MC Tests | 3.81 | 1.84 | | | | | | | | |
| 9. Experience with CR Tests | 3.03 | 1.16 | — | | | | | | | |
| 10. Face Validity | 2.79/2.70 | 0.70/0.75 | -06 | (79) | | | | | | |
| 11. Fairness | 3.12/3.12 | 0.57/0.52 | -09 | 21/35 | (64) | | | | | |
| 12. Predictive Validity | 2.14/2.12 | 0.68/0.73 | 02 | 43/54 | 46/50 | (82) | | | | |
| 13. Belief in Tests | 2.60/2.66 | 0.73/0.68 | -05 | 34/54 | 45/56 | 57/62 | (70) | | | |
| 14. Stereotype Threat | 2.77/2.74 | 0.48/0.52 | 17 | -09/-09 | -32/-33 | -23/-18 | -37/-36 | (47) | | |
| 15. Self–Efficacy | 3.28/3.28 | 0.78/0.79 | -15 | -04/20 | 40/38 | 19/31 | 34/38 | -58/-60 | (83) | |
| 16. Motivation | 3.92/3.93 | 0.71/0.67 | -12 | 00/30 | 22/27 | 03/25 | 17/28 | -16/-26 | 41/48 | (91) |
| 17. Level of Effort | 2.69 | 0.95 | -11 | 03 | 24 | 13 | 24 | -28 | 44 | 48 |
| 18. Anxiety | 2.92/2.89 | 0.68/0.72 | 10 | 03/-20 | -35/-40 | -23/-31 | -31/-40 | 67/65 | -70/-81 | -09/-35 |
| 19. College GPA | 5.44 | 1.11 | 00 | -05 | 08 | 02 | 10 | -12 | 19 | 10 |

Table 3 Continued

| Variable | *M* | *SD* | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|
| 20. High School GPA | 6.38 | 0.84 | 05 | -12 | 08 | 01 | 10 | -22 | 11 | 07 |
| 21. High School Rank | 4.58 | 1.86 | 04 | -08 | 03 | -03 | 05 | -17 | 09 | 00 |
| 22. SAT | 1096.90 | 167.07 | 12 | -02 | 26 | 16 | 22 | -32 | 30 | -01 |
| 23. Test Performance[D] | 12.21 | 4.36 | 08 | 00 | 21 | 12 | 15 | -32 | 28 | 05 |

Table 3 Continued

| Variable | *M* | *SD* | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|
| 1. MC[A] Test | 13.40 | 4.18 | | | | | | | |
| 2. CR[B] Test | 11.11 | 4.24 | | | | | | | |
| 3. Race[C] | — | — | | | | | | | |
| 4. Cognitive Ability | 7.44 | 2.40 | | | | | | | |
| 5. Reading Ability | 17.04 | 2.99 | | | | | | | |
| 6. Test–taking Skills | 8.60 | 3.10 | | | | | | | |
| 7. Standardized Test Coaching | 5.26 | 1.04 | | | | | | | |
| 8. Experience with MC Tests | 3.81 | 1.84 | | | | | | | |
| 9. Experience with CR Tests | 3.03 | 1.16 | | | | | | | |
| 10. Face Validity | 2.79/2.70 | 0.70/0.75 | | | | | | | |
| 11. Fairness | 3.12/3.12 | 0.570/.52 | | | | | | | |
| 12. Predictive Validity | 2.14/2.12 | 0.68/0.73 | | | | | | | |
| 13. Belief in Tests | 2.60/2.66 | 0.73/0.68 | | | | | | | |
| 14. Stereotype Threat | 2.77/2.74 | 0.48/0.52 | | | | | | | |
| 15. Self–Efficacy | 3.28/3.28 | 0.78/0.79 | | | | | | | |
| 16. Motivation | 3.92/3.93 | 0.71/0.67 | | | | | | | |
| 17. Level of Effort | 2.69 | 0.95 | — | | | | | | |
| 18. Anxiety | 2.92/2.89 | 0.68/0.72 | -31 | (86) | | | | | |
| 19. College GPA | 5.44 | 1.11 | 08 | -23 | (99) | | | | |

Table 3 Continued

| Variable | *M* | *SD* | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|
| 20. High School GPA | 6.38 | 0.84 | 11 | -19 | 15 | (94) | | | |
| 21. High School Rank | 4.58 | 1.86 | 09 | -15 | 07 | 52 | (99) | | |
| 22. SAT | 1096.90 | 167.07 | 16 | -35 | 17 | 42 | 32 | (95) | |
| 23. Test Performance[D] | 12.21 | 4.36 | 20 | -31 | 15 | 31 | 24 | 56 | (80) |

NOTE: [A]Multiple–choice test performance ($n$ = 222). [B]Constructed response test performance ($n$ = 242). [C]Race is coded 1 = White, 0 = African American. [D]Test performance across both formats. Decimals have been removed from the correlations. Correlations including test–taking skills are computed using $n$ = 356. The first *M*, *SD*, and correlation for variables 10–16 and 18 is with the multiple–choice test condition ($n$ = 222) and the second correlation is with the constructed response test condition ($n$ = 242). Reliabilities are in the diagonal (an odd–even, split–half reliability with a Spearman–Brown correction is reported for the multiple–choice test, constructed response test, cognitive ability, and test performance; coefficient alpha is reported for reading ability, test–taking skills, fairness, predictive validity, belief in tests, stereotype threat, self–efficacy, test–taking motivation, and test–taking anxiety; test–retest is reported for college GPA, high school GPA, high school rank, and SAT scores). If $r$ = |10| to |12| then $p < .05$; if $r$ = |13| to |16| then $p < .01$; if $r >$ |16| then $p < .001$.

Table 4

*Descriptive Statistics and Correlations for Hypothesis–Related Study Variables*

| Variable | *M* | *SD* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. MC[A] Test | 13.40 | 4.18 | (83) | | | | | | | |
| 2. CR[B] Test | 11.11 | 4.24 | — | (78) | | | | | | |
| 3. Race[C] | — | — | 56 | 43 | — | | | | | |
| 4. Reading Ability | 17.04 | 2.99 | 59 | 50 | 44 | (80) | | | | |
| 5. Test–taking Skills | 8.60 | 3.10 | 33 | 36 | 21 | 33 | (57) | | | |
| 6. Face Validity | 2.79/2.70 | 0.70/0.75 | -04 | 00 | -10 | -05 | -04 | (79) | | |
| 7. Fairness | 3.12/3.12 | 0.57/0.52 | 30 | 17 | 12 | 19 | 11 | 21/35 | (64) | |
| 8. Predictive Validity | 2.14/2.12 | 0.68/0.73 | 13 | 12 | 00 | 04 | 04 | 43/54 | 46/50 | (82) |
| 9. Belief in Tests | 2.60/2.66 | 0.73/0.68 | 15 | 19 | 00 | 11 | 01 | 34/54 | 45/56 | 57/62 |
| 10. Stereotype Threat | 2.77/2.74 | 0.48/0.52 | -27 | -41 | -08 | -22 | -11 | -09/-09 | -32/-33 | -23/-18 |
| 11. Self–Efficacy | 3.28/3.28 | 0.78/0.79 | 22 | 35 | -12 | 17 | 16 | -04/20 | 40/38 | 19/31 |
| 12. Motivation | 3.92/3.93 | 0.71/0.67 | 02 | 09 | -19 | 01 | -05 | 00/30 | 22/27 | 03/25 |
| 13. Anxiety | 2.92/2.89 | 0.68/0.72 | -28 | -37 | 06 | -20 | -17 | 03/-20 | -35/-40 | -23/-31 |
| 14. College GPA | 5.44 | 1.11 | 13 | 20 | 05 | 11 | 07 | -05 | 08 | 02 |
| 15. High School GPA | 6.38 | 0.84 | 36 | 46 | 27 | 28 | 11 | -12 | 08 | 01 |
| 16. High School Rank | 4.58 | 1.86 | 30 | 27 | 21 | 23 | 11 | -08 | 03 | -03 |
| 17. SAT | 1096.90 | 167.07 | 62 | 65 | 51 | 53 | 36 | -02 | 26 | 16 |
| 18. Test Performance[D] | 12.21 | 4.36 | — | — | 43 | 50 | 32 | 00 | 21 | 12 |

Table 4 Continued

| Variable | *M* | *SD* | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. MC[A] Test | 13.40 | 4.18 | | | | | | | | |
| 2. CR[B] Test | 11.11 | 4.24 | | | | | | | | |
| 3. Race[C] | — | — | | | | | | | | |
| 4. Reading Ability | 17.04 | 2.99 | | | | | | | | |
| 5. Test–taking Skills | 8.60 | 3.10 | | | | | | | | |
| 6. Face Validity | 2.79/2.70 | 0.70/0.75 | | | | | | | | |
| 7. Fairness | 3.12/3.12 | 0.57/0.52 | | | | | | | | |
| 8. Predictive Validity | 2.14/2.12 | 0.68/0.73 | | | | | | | | |
| 9. Belief in Tests | 2.60/2.66 | 0.73/0.68 | (70) | | | | | | | |
| 10. Stereotype Threat | 2.77/2.74 | 0.48/0.52 | -37/-36 | (47) | | | | | | |
| 11. Self–Efficacy | 3.28/3.28 | 0.78/0.79 | 34/38 | -58/-60 | (83) | | | | | |
| 12. Motivation | 3.92/3.93 | 0.71/0.67 | 17/28 | -16/-26 | 41/48 | (91) | | | | |
| 13. Anxiety | 2.92/2.89 | 0.68/0.72 | -31/-40 | 67/65 | -70/-81 | -09/-35 | (86) | | | |
| 14. College GPA | 5.44 | 1.11 | 10 | -12 | 19 | 10 | -23 | (99) | | |
| 15. High School GPA | 6.38 | 0.84 | 10 | -22 | 11 | 07 | -19 | 15 | (94) | |
| 16. High School Rank | 4.58 | 1.86 | 05 | -17 | 09 | 00 | -15 | 07 | 52 | (99) |
| 17. SAT | 1096.90 | 167.07 | 22 | -32 | 30 | -01 | -35 | 17 | 42 | 32 |
| 18. Test Performance[D] | 12.21 | 4.36 | 15 | -32 | 28 | 05 | -31 | 15 | 31 | 24 |

Table 4 Continued

| Variable | *M* | *SD* | 17 | 18 |
|---|---|---|---|---|
| 1. MC[A] Test | 13.40 | 4.18 | | |
| 2. CR[B] Test | 11.11 | 4.24 | | |
| 3. Race[C] | — | — | | |
| 4. Reading Ability | 17.04 | 2.99 | | |
| 5. Test–taking Skills | 8.60 | 3.10 | | |
| 6. Face Validity | 2.79/2.70 | 0.70/0.75 | | |
| 7. Fairness | 3.12/3.12 | 0.57/0.52 | | |
| 8. Predictive Validity | 2.14/2.12 | 0.68/0.73 | | |
| 9. Belief in Tests | 2.60/2.66 | 0.73/0.68 | | |
| 10. Stereotype Threat | 2.77/2.74 | 0.48/0.52 | | |
| 11. Self–Efficacy | 3.28/3.28 | 0.78/0.79 | | |
| 12. Motivation | 3.92/3.93 | 0.71/0.67 | | |
| 13. Anxiety | 2.92/2.89 | 0.68/0.72 | | |
| 14. College GPA | 5.44 | 1.11 | | |
| 15. High School GPA | 6.38 | 0.84 | | |
| 16. High School Rank | 4.58 | 1.86 | | |
| 17. SAT | 1096.90 | 167.07 | (95) | |
| 18. Test Performance[D] | 12.21 | 4.36 | 56 | (80) |

Table 4 Continued

NOTE: [A]Multiple–choice test performance ($n$ = 222). [B]Constructed response test performance ($n$ = 242). [C]Race is coded 1 = White, 0 = African American. [D]Test performance across both formats. Decimals have been removed from the correlations. Correlations including test–taking skills are computed using $n$ = 356. The first $M$, $SD$, and correlation for variables 6–13 is with the multiple–choice test condition ($n$ = 222) and the second correlation is with the constructed response test condition ($n$ = 242). Reliabilities are in the diagonal (an odd–even, split–half reliability with a Spearman–Brown correction is reported for multiple–choice test, constructed response test, and test performance; coefficient alpha is reported for reading ability, test–taking skills, fairness, predictive validity, belief in tests, stereotype threat, self–efficacy, test–taking motivation, and test–taking anxiety; test–retest is reported for college GPA, high school GPA, high school rank, and SAT scores). If $r$ = |10| to |12| then $p < .05$; if $r$ = |13| to |16| then $p < .01$; if $r >$ |16| then $p < .001$.

experience with taking multiple–choice tests and test–taking skills ($r$ = .03, *ns*, 95% CI = -.07 to .13) and between experience with taking constructed response tests and test–taking skills ($r$ = -.02, *ns*, 95% CI = -.12 to .08).

### *Level of Effort in the Study*

As a check on participants' level of effort exerted in the present study, one item was administered at the end of the protocol to measure participants' motivation to perform well on all the tests. The overall mean for the level of effort exerted in the protocol ($M$ = 2.68, $SD$ = .95), indicated that participants were "quite a bit motivated to do their best." There was no difference ($t$ [461] = 1.29, *ns*; $d$ = .13) in reported level of effort between African Americans and Whites.

### *Subgroup Differences on the Multiple–Choice and Constructed Response Tests*

The results presented in Table 4 show that performance was higher on the multiple–choice test than the constructed response test, $d$ = .52, $t$(463) = 5.84, $p$ < .001. To test Hypothesis 1, a 2 (test format) × 2 (race) analysis of variance (ANOVA) was conducted. Both the main effect for test format ($F$ [1, 460] = 71.37, $p$ < .001, $\omega^2$ = .11) and race ($F$ [1, 460] = 119.21, $p$ < .001, $\omega^2$ = .18) were significant, but not the interaction ($F$ [1, 460] = 1.57, *ns*). Therefore, contrary to Hypothesis 1, although African American–White subgroup differences on the constructed response test appeared to be smaller ($d$ = .98) than differences on the multiple–choice test ($d$ = 1.33), the reduction was not statistically significant. The above results are further illustrated in Figure 2, which presents subgroup differences on the multiple–choice and constructed response tests.

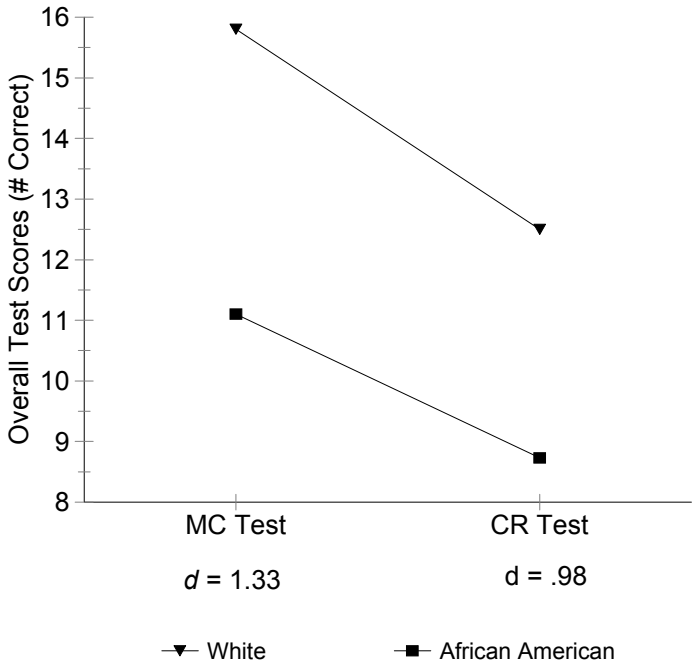


*Figure 2*. Subgroup differences in multiple–choice and constructed response test performance.

*Supplementary Analyses for Hypothesis 1*

The 12–item short form of the APM was administered in the present study to statistically control for any preexisting differences between groups on ability level. There were minimal differences in APM scores for the multiple–choice ($M$ = 7.13, $SD$ = 2.55) and constructed response test ($M$ = 7.72, $SD$ = 2.24) conditions ($t$ [262] = 2.68, $p$ < .05, $d$ = .25). In contrast, subgroup differences on cognitive ability were quite large ($t$ [262] = 9.52, $p$ < .001, $d$ = .90) with Whites ($M$ = 8.30, $SD$ = 2.15) scoring higher than African Americans ($M$ = 6.33, $SD$ = 2.26). Given preexisting differences in cognitive ability for race and the two testing conditions, a 2 (test format) × 2 (race) analysis of covariance (ANCOVA) was conducted using cognitive ability as a covariate. Again, both the main effect for test format ($F$ [1, 459] = 102.52, $p$ < .001, $\omega^2$ = .06) and race ($F$

[1, 459] = 48.66, $p < .001$, $\omega^2 = .12$) were significant, but not the interaction ($F$ [1, 459] = .81, *ns*). Therefore, using cognitive ability as a covariate to statistically remove preexisting differences, subgroup differences did not differ as a function of test format. Thus, although the pattern of results were in the predicted direction, Hypothesis 1 was not supported.

The primary concern in comparing test formats that purportedly measure the same content/construct domain is that the test formats themselves may introduce different construct irrelevant variance. Indeed, the primary criticism of multiple–choice tests is that this format measures constructs that are unrelated to the content or a performance criterion. Traub (1993) indicates that some traits or constructs may not be measured similarly by two different test formats. For example, he indicated that scores on writing and word knowledge tests may be affected differentially by format while reading comprehension and most quantitative skills may not yield format effects. The multiple–choice and constructed response tests used in the present study measured two very different constructs—math and science reasoning. Given Traub's (1993) interpretation of the literature it seems likely that performance may be impacted by the construct being measured in the present study (i.e., math vs. science reasoning). Therefore, Hypothesis 1 was tested for scores on the math and science reasoning sections only.

*Math Section*. Performance was higher on the math section of the multiple–choice test ($M = 7.04$, $SD = 2.14$) than on the math section of the constructed response test, ($M = 5.31$, $SD = 2.44$; $d = .75$; $t[463] = 8.11$, $p < .001$). To test Hypothesis 1 for scores on the math section only, a 2 (test format) × 2 (race) ANOVA

was conducted. Both the main effect for test format ($F$ [1, 460] = 94.92, $p < .001$, $\omega^2 = .16$) and race ($F$ [1, 460] = 45.57, $p < .001$, $\omega^2 = .07$) were significant, but not the interaction ($F$ [1, 460] = .01, *ns*). Therefore, similar to the results obtained for overall test performance, although African American–White subgroup differences on the math section of the constructed response test appeared to be smaller (African American $M$ = 4.25, $SD$ = 2.46; White $M$ = 5.93, $SD$ = 2.22; $d$ = .73) than differences on the math section of the multiple–choice test (African American $M$ = 6.24, $SD$ = 2.22; White $M$ = 7.89, $SD$ = 1.69; $d$ = .83), the reduction was not statistically significant. The above results are further illustrated in Figure 3, which presents subgroup differences on the math sections of the multiple–choice and constructed response tests.

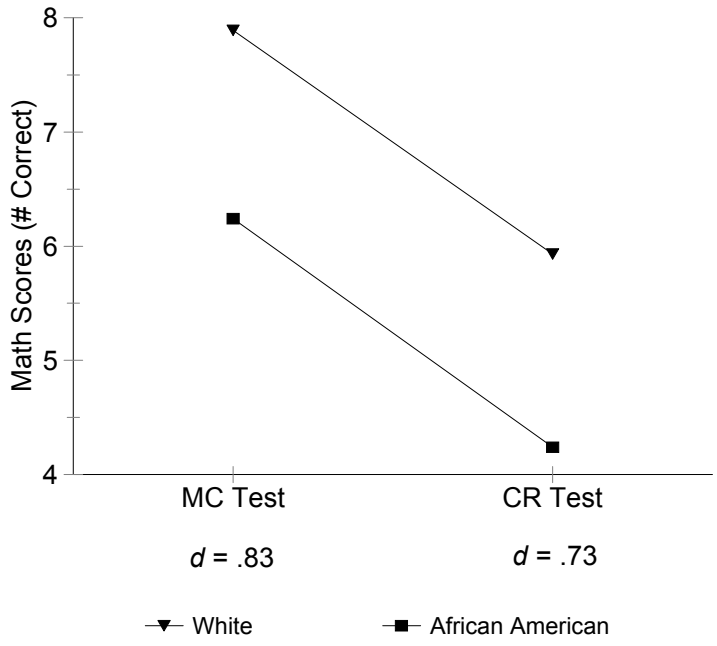


*Figure 3*. Subgroup differences in test performance on the math sections of the multiple–choice and constructed response tests.

*Science Reasoning Section*. Performance was higher on the science reasoning section of the multiple–choice test ($M = 6.36$, $SD = 2.60$) than on the science reasoning section of the constructed response test, ($M = 5.80$, $SD = 2.40$; $d = .22$; $t[463] = 2.42$, $p < .05$). To test Hypothesis 1 for scores on the science reasoning section only, a 2 (test format) × 2 (race) ANOVA was conducted. The main effects for test format ($F$ [1, 460] = 21.74, $p < .001$, $\omega^2 = .03$) and race ($F$ [1, 460] = 141.05, $p < .001$, $\omega^2 = .22$) and the interaction ($F$ [1, 460] = 4.97, $p < .05$, $\omega^2 = .01$) were significant. Contrary to the results for overall test performance and performance on the math section only, African American–White subgroup differences on the science reasoning section of the constructed response test were significantly smaller (African American $M = 4.48$, $SD = 2.53$; White $M = 6.56$, $SD = 1.95$; $d = .96$) than differences on the science reasoning section of the multiple–choice test (African American $M = 4.92$, $SD = 2.34$; White $M = 7.91$, $SD = 1.87$; $d = 1.41$). Therefore, Hypothesis 1 was supported for analyses using only test performance on the science reasoning section as the dependent variable. The above results are further illustrated in Figure 4, which presents subgroup differences on the science reasoning sections of the multiple–choice and constructed response tests. It is important to note that this effect is due to differentiated decrements and not improvements in test performance.

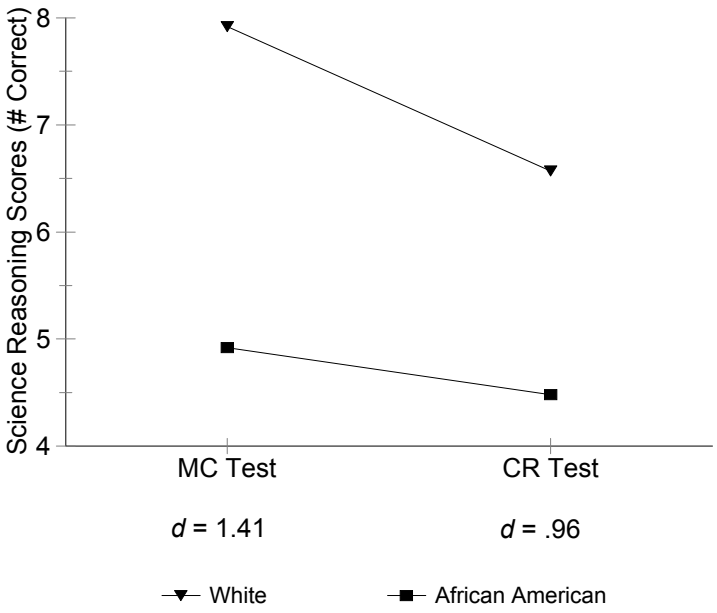


*Figure 4*. Subgroup differences in test performance on the science reasoning sections of the multiple–choice and constructed response tests.

### *Reading Ability*

The relationships between race, reading ability and multiple–choice and constructed response test performance posited in Hypotheses 2a–2d are predicated on the assumption that the reading load of the multiple–choice test was higher than a stem–equivalent constructed response test. The results of a readability analysis which are presented in Table 5, show that the constructed response test consisted of 27% fewer words and 46% fewer sentences than the multiple–choice test. However, their reading grade levels were comparable. Therefore, the reading load on the multiple–choice test was higher than the reading load on the constructed response test.

Table 5
*Readability Analysis for the Multiple–Choice and Constructed Response Test Formats*

| | Test Items | | |
|---|---|---|---|
| | Multiple–Choice | | Constructed Response |
| Statistic | Item Stem and Alternatives | Item Stem Only | Item Stem Only[A] |
| Number of items | 20 | 20 | 20 |
| Number of words | 1,438 | 1,048 | 1,058 |
| Number of sentences | 421 | 229 | 250 |
| Average word per sentence | 3 | 5 | 4 |
| Average word length | 5 | 5 | 5 |
| Grade level (Flesch–Kincaid) | 9.12 | 9.35 | 9.10 |

Note. [A]Constructed response format does not have any alternatives.

Hypotheses 2a–2d were tested by running a series of regressions (see Table 6). Hypothesis 2a predicted a significant, positive relationship between reading ability and test performance. In support of Hypothesis 2a, there was a significant relationship between reading ability and test performance ($R^2 = .25$, $p < .001$). Next, Hypothesis 2b predicted a significant relationship between reading ability and test format such that the reading ability/multiple–choice test performance relationship would be stronger than the reading ability/constructed response test performance relationship. To test this hypothesis, reading ability was regressed on the test format × test performance interaction term which was not significant ($R^2 = .01$, *ns*). Consistent with this, the difference between the reading ability/multiple–choice test performance correlation ($r = .59$, $p < .001$, 95% CI = .50 to .67) and reading ability/constructed response test

performance correlation ($r$ = .50, $p$ < .001, 95% CI = .40 to .59) was not significant ($z$ = 1.37, *ns*). Therefore, the relationship between reading ability and test performance did not differ as a function of test format. Hypothesis 2c predicted that there would be significant subgroup differences in reading ability. Consistent with this prediction, there was a significant relationship between race and reading ability ($R^2$ = .19, $p$ < .001) indicating that there were large, subgroup differences on reading ability ($d$ = .98). Specifically, Whites ($M$ = 18.16, $SD$ = 1.94) scored higher on reading ability than African Americans ($M$ = 15.61, $SD$ = 3.45).

Table 6
*Results of Regressions for Reading Ability (Hypotheses 2a–2d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 2a | Test Performance | Reading Ability | .50* | .25* | |
| 2b | Reading Ability | Test Format × Test Performance | .08 | .01 | |
| 2c | Reading Ability | Race | .44* | .19* | |
| 2d | Test Performance | Race | .43* | **.18*** | |
| | Reading Ability | Race | .44* | .19* | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Reading Ability | .50* | .25* | |
| Step 2 | Test Performance | Reading Ability | .38* | .25* | |
| | | Race | .26* | .30* | **.05*** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p$ < .001.

Finally, Hypothesis 2d predicted that reading ability would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test. The hypothesized mediation effects specified in Hypothesis 2d, were assessed in accordance with standards outlined by Baron and Kenny (1986) who specify three conditions that must be met to infer mediation—(a) race (independent variable) must be related to test performance (dependent variable), (b) race must be related to reading ability (mediator variable), and (c) when the independent variable and mediator are considered simultaneously, the direct relationship between the independent and dependent variable should show a significant decrease (partial mediation). The results presented in Table 6 indicate that all the criteria for mediation were met. In addition, they show a significant decrease in the unique variance explained by race ($\Delta R^2$ = .05 vs. $R^2$ = .18) after controlling for reading ability, demonstrating support for partial mediation. Finally, Sobel's (1982) test for the indirect effect of race and test performance through reading ability was significant (Sobel = 6.40, $p < .01$). Given that reading ability was a partial mediator, a 2 (test format) × 2 (race) ANCOVA was computed with reading ability used as a covariate to further test Hypothesis 2d. Again, the main effects for test format ($F$ [1, 459] = 91.92, $p < .001$, $\omega^2 = .12$) and race ($F$ [1, 459] = 43.33, $p < .001$, $\omega^2 = .05$) were significant, but the interaction was not ($F$ [1, 459] = .17, *ns*).

In summary, there was support for Hypotheses 2a, 2c, and 2d in that reading ability was related to test performance, there were subgroup differences on reading ability, and reading ability partially mediated the relationship between race and test

performance. However, the strength of mediation was not sufficient to reduce subgroup differences on the multiple–choice test to the levels of subgroup differences observed on the constructed response test.

*Test–Taking Skills*

A number of regressions were run to test Hypotheses 3a–3d (see Table 7). Hypothesis 3a predicted a significant, positive relationship between test–taking skills and test performance. In support of Hypothesis 3a, there was a significant relationship between test–taking skills and test performance ($R^2$ = .10, $p$ < .001). Next, Hypothesis 3b predicted a significant relationship between test–taking skills and test format such that the test–taking skills/multiple–choice test performance relationship would be stronger than the test–taking skills/constructed response test relationship. To test this hypothesis, test–taking skills was regressed on the test format × test performance interaction term which was not significant ($R^2$ = .00, *ns*). Consistent with this, the difference between the test–taking skills/multiple–choice test performance correlation ($r$ = .33, $p$ < .001, 95% CI = .18 to .46) and test–taking skills/constructed response test performance correlation ($r$ = .36, $p$ < .001, 95% CI = .23 to .48) was not significant ($z$ = .36, *ns*). Therefore, the relationship between test–taking skills and test performance did not differ as a function of test format. Hypothesis 3c predicted that there would be significant subgroup differences on test–taking skills. Consistent with this prediction, there was a significant relationship between race and test–taking skills ($R^2$ = .05, $p$ < .001) indicating that there were moderate subgroup differences on test–taking skills ($d$ = .49). Specifically, Whites scored higher ($M$ = 9.01, $SD$ = 3.14) on test–taking skills than African Americans ($M$ = 7.52, $SD$ = 2.74).

Table 7
*Results of Regressions for Test–Taking Skills (Hypotheses 3a–3d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 3a | Test Performance | Test–Taking Skills | .32* | .10* | |
| 3b | Test–Taking Skills | Test Format × Test Performance | .07 | .00 | |
| 3c | Test–Taking Skills | Race | .21* | .05* | |
| 3d | Test Performance | Race | .31* | **.09*** | |
| | Test–Taking Skills | Race | .21* | .05* | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Test–Taking Skills | .32* | .10* | |
| Step 2 | Test Performance | Test–Taking Skills | .27* | .10* | |
| | | Race | .25* | .16* | **.06*** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .001$.

Finally, Hypothesis 3d predicted that test–taking skills would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test. The results presented in Table 7 indicate that all the criteria for mediation were met. In addition, they show a significant decrease in the unique variance explained by race ($\Delta R^2 = .06$ vs. $R^2 = .09$) after controlling for test–taking skills, demonstrating support for partial mediation. Finally, Sobel's (1982) test for the indirect effect of race and test performance through test–taking skills was significant (Sobel = 4.02, $p < .001$). Given that test–taking skills was a partial mediator, a 2 (test format) × 2 (race) ANCOVA was computed with test–taking skills used as a covariate

to further test Hypothesis 3d. The main effects for test format ($F$ [1, 351] = 76.64, $p$ < .001, $\omega^2$ = .15) and race ($F$ [1, 351] = 30.01, $p$ < .001, $\omega^2$ = .06) were significant, but the interaction was not ($F$ [1, 351] = .00, *ns*).

In summary, there was support for Hypotheses 3a, 3c, and 3d in that test–taking skills was related to test performance, there were subgroup differences on test–taking skills, and test–taking skills partially mediated the relationship between race and test performance. However, the strength of mediation was not sufficient to reduce subgroup differences on the multiple–choice test to the levels of subgroup differences observed on the constructed response test.

*Test Perceptions*

As shown in Table 4, test perceptions were similar across the two test formats. For example, standardized differences on test perceptions between the two conditions were minimal and ranged from $d$ = .00 to .12. The largest differences were for face validity ($d$ = .12), belief in tests ($d$ = .09) and stereotype threat ($d$ = .06). Contrary to expectations, the multiple–choice test was seen as being more face valid than the constructed response test. However, reported stereotype threat was higher in the multiple–choice than the constructed response test condition. Belief in tests was the only construct that was not test specific as it refers to the strength of the belief that tests in general are valid indicators of job performance. Nevertheless, it is entirely likely that responses to this measure were influenced by test format as participants in the constructed response test condition reported higher levels of belief in tests. Therefore, where relevant, analyses related to belief in tests were conducted by test condition, similar to the other test perceptions.

*Face Validity*

Hypotheses 4a–4d were tested by running a series of regressions (see Table 8). Hypothesis 4a predicted a significant, positive relationship between face validity and test performance. Contrary to Hypothesis 4a, there was no relationship between face validity and test performance ($R^2$ = .00, *ns*). Next, Hypothesis 4b predicted a significant relationship between face validity and test format such that the face validity/multiple–choice test performance relationship would be stronger than the face validity/constructed response test performance relationship. To test this hypothesis, face validity was regressed on the test format × test performance interaction term which was not significant ($R^2$ = .00, *ns*). Therefore, the relationship between face validity and test performance did not differ as a function of test format. Hypothesis 4c predicted that there would be significant subgroup differences on face validity. Given that ratings of face validity were made in reaction to having been exposed to one of the two testing conditions, the interaction between race and test format was regressed onto face validity. The results in Table 8 show that there was a significant relationship between race and face validity ($R^2$ = .01, $p < .05$) indicating that there were small subgroup differences on face validity ($d$ = .21). Specifically, African Americans reported higher levels of face validity ($M$ = 2.82, $SD$ = .73) than Whites ($M$ = 2.67, $SD$ = .72). However, the relationship between race and face validity did not differ as a function of test format ($\beta$ = .15, *ns*).

Finally, Hypothesis 4d predicted that face validity would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the

constructed response test. The results presented in Table 8 indicate that all the criteria for mediation were met. However, they show that there was no decrease in the unique variance explained by race ($\Delta R^2$ = .18 vs. $R^2$ = .18) after controlling for face validity, so there was no support for partial mediation.

Table 8
*Results of Regressions for Face Validity (Hypotheses 4a–4d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 4a | Test Performance | Face Validity | .00 | .00 | |
| 4b | Face Validity | Test Format × Test Performance | .05 | .00 | |
| 4c | Face Validity | Race | -.18** | | |
| | Face Validity | Test Format | -.06 | | |
| | Face Validity | Test Format × Race | .15 | .02* | |
| 4d | Test Performance | Race | .43*** | **.18*** | |
| | Face Validity | Race | -.10* | .01* | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Face Validity | .00 | .00 | |
| Step 2 | Test Performance | Face Validity | .04 | .00 | |
| | | Race | .43*** | .18*** | **.18*** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .05$, ** $p < .01$, ***$p < .001$.

In summary, support was obtained for only Hypothesis 4c in that there were significant subgroup differences in face validity. There was no support for Hypotheses 4a, 4b and 4d. Therefore, face validity was not related to test performance, did not

differ as a function of test format, and did not appear to mediate the relationship between race and test performance.

*Fairness*

A number of regressions were run to test Hypotheses 5a–5d (see Table 9). Hypothesis 5a predicted a significant, positive relationship between fairness and test performance. Consistent with Hypothesis 5a, there was a significant relationship between fairness and test performance ($R^2$ = .04, $p$ < .001). Next, Hypothesis 5b predicted a significant relationship between fairness and test format such that the fairness/multiple–choice test performance relationship would be stronger than the fairness/constructed response test performance relationship. To test this hypothesis, fairness was regressed on the test format × test performance interaction term which was not significant ($R^2$ = .00, *ns*). Consistent with this, the difference between the fairness/multiple–choice test performance correlation ($r$ = .30, $p$ < .001, 95% CI = .18 to .42) and the fairness/constructed response test performance correlation ($r$ = .17, $p$ < .001, 95% CI = .04 to .29) was not significant ($z$ = 1.47, *ns*). Therefore, the relationship between fairness and test performance did not differ as a function of test format. Hypothesis 5c predicted that there would be significant subgroup differences on fairness. Given that ratings of fairness were made in reaction to having been exposed to one of the two testing conditions, the interaction between race and test format was regressed onto fairness. The results in Table 9 show that there was a significant relationship between race and fairness as a function of test format ($\beta$ = .26, $p$ < .001). Specifically, Whites ($M$ = 3.26, $SD$ = .49) reported higher levels of fairness than African Americans ($M$ = 2.98, $SD$ = .61; $d$ = .54) on the multiple–choice test and

African Americans ($M$ = 3.22, $SD$ = .52) reported higher levels of fairness than Whites ($M$ = 3.18, $SD$ = .53; $d$ = .09) on the constructed response test. The above results are further illustrated in Figure 5, which presents subgroup differences in reported fairness in the multiple–choice and constructed response test conditions.

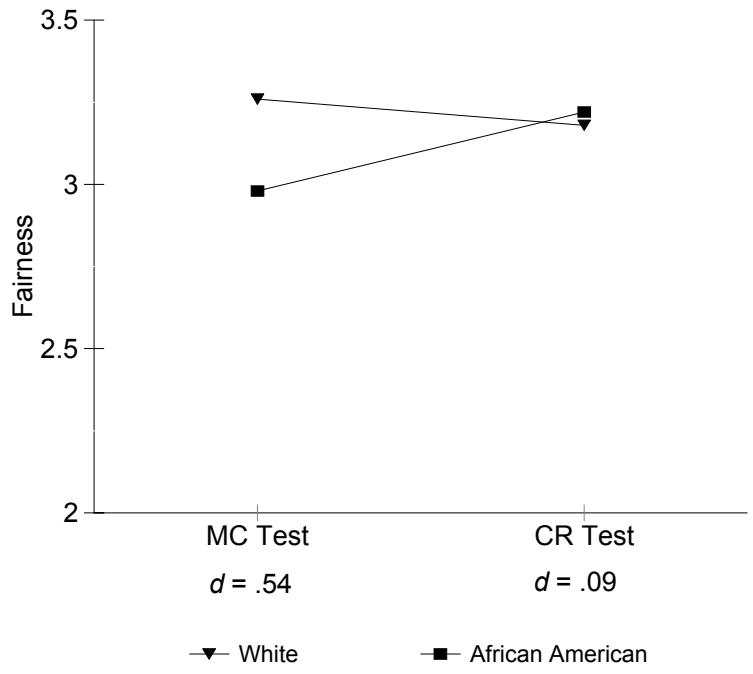


*Figure 5*. Subgroup differences in fairness in the multiple–choice and constructed response test conditions.

Finally, Hypothesis 5d predicted that fairness would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test. The results in Table 9 indicate that all the criteria for mediation were met. In addition, they show that there was a significant decrease in the unique variance explained by race ($\Delta R^2$ = .17 vs. $R^2$ = .18) after controlling for fairness,

so there was support for partial mediation. Finally, Sobel's (1982) test for the indirect effect of race and test performance through fairness was significant (Sobel = 3.55, $p < .001$). Given that fairness was a partial mediator, a 2 (test format) × 2 (race) ANCOVA was computed with fairness used as a covariate to further test Hypothesis 5d. The main effects for test format ($F$ [1, 459] = 78.80, $p < .001$, $\omega^2 = .11$) and race ($F$ [1, 459] = 111.59, $p < .001$, $\omega^2 = .16$) were significant, but the interaction was not ($F$ [1, 351] = .33, *ns*).

Table 9
*Results of Regressions for Fairness (Hypotheses 5a–5d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 5a | Test Performance | Fairness | .21** | .04** | |
| 5b | Fairness | Test Format × Test Performance | .02 | .00 | |
| 5c | Fairness | Race | -.04 | | |
| | Fairness | Test Format | -.23** | | |
| | Fairness | Test Format × Race | .26** | .04** | |
| 5d | Test Performance | Race | .43** | **.18**** | |
| | Fairness | Race | .12* | .01* | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Fairness | .21** | .04** | |
| Step 2 | Test Performance | Fairness | .16** | .04** | |
| | | Race | .41** | .21** | **.17**** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .01$, **$p < .001$.

In summary, support was obtained for Hypotheses 5a, 5c, and 5d in that fairness was related to test performance, there were significant subgroup differences in fairness as a function of test format, and fairness mediated the relationship between race and test performance. However, the strength of the mediation was not sufficient to reduce subgroup differences on the multiple–choice test to the levels of subgroup differences observed on the constructed response test.

*Perceived Predictive Validity*

Hypotheses 6a–6d were tested by running a series of regressions (see Table 10). Hypothesis 6a predicted a significant, positive relationship between perceived predictive validity and test performance. Consistent with Hypothesis 6a, there was a significant relationship between perceived predictive validity and test performance ($R^2$ = .02, $p$ < .01). Next, Hypothesis 6b predicted a significant relationship between perceived predictive validity and test format such that the perceived predictive validity/multiple–choice test performance relationship would be stronger than the perceived predictive validity/constructed response test performance relationship. To test this hypothesis, perceived predictive validity was regressed on the test format × test performance interaction term which was not significant ($R^2$ = .00, *ns*). Consistent with this, the difference between the perceived predictive validity/multiple–choice test performance correlation ($r$ = .13, $p$ < .01, 95% CI = .00 to .26) and the perceived predictive validity/constructed response test performance correlation ($r$ = .12, $p$ < .05, 95% CI = -.01 to .24) was not significant ($z$ = .11, *ns*). Therefore, the relationship between perceived predictive validity and test performance did not differ as a function of test format. Hypothesis 6c predicted that there would be significant subgroup

differences on perceived predictive validity. Given that ratings of perceived predictive validity were made in reaction to having been exposed to one of the two testing conditions, the interaction between race and test format was regressed onto perceived predictive validity. The results in Table 10 show that there was a significant relationship between race and perceived predictive validity as a function of test format ($\beta = .18$, $p < .05$). Specifically, Whites ($M = 2.22$, $SD = .63$) reported higher levels of perceived predictive validity than African Americans ($M = 2.06$, $SD = .71$; $d = .24$) on the multiple–choice test and African Americans ($M = 2.21$, $SD = .73$; $d = .19$) reported higher levels of face validity than Whites ($M = 2.07$, $SD = .73$) on the constructed response test. The above results are further illustrated in Figure 6, which presents subgroup differences in perceived predictive validity in the multiple–choice and constructed response test conditions.

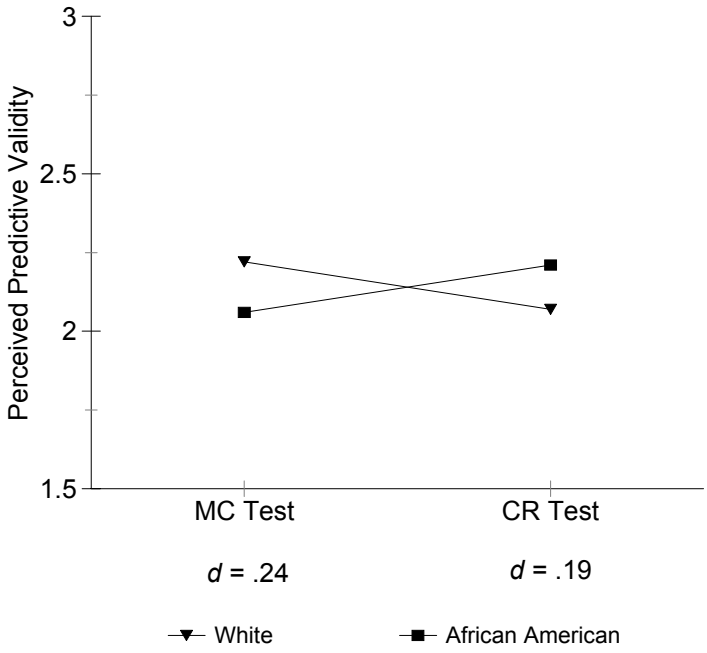


*Figure 6.* Subgroup differences in perceived predictive validity in the multiple–choice and constructed response test conditions.

Finally, Hypothesis 6d posited that perceived predictive validity would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test. The results presented in Table 10 indicate that the relationship between race and perceived predictive validity was not significant, so the criteria for mediation were not met.

Table 10
*Results of Regressions for Perceived Predictive Validity (Hypotheses 6a–6d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 6a | Test Performance | Predictive[A] | .12** | .02** | |
| 6b | Predictive | Test Format × Test Performance | .05 | .00 | |
| 6c | Predictive | Race | -.10 | | |
| | Predictive | Test Format | -.11 | | |
| | Predictive | Test Format × Race | .18* | .01 | |
| 6d | Test Performance | Race | .43*** | **.18*** ** | |
| | Predictive | Race | .00 | .00 | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Predictive | .12** | .02** | |
| Step 2 | Test Performance | Predictive | .12** | .02** | |
| | | Race | .43*** | .20*** | **.18**** |

*Note.* [A]Perceived predictive validity. Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .05$, **$p < .01$, ***$p < .001$.

In summary, support was obtained for only Hypotheses 6a and 6c in that perceived predictive validity was related to test performance and there were significant subgroup differences as a function of test format. However, there was no support for Hypotheses 6b and 6d. As such, perceived predictive validity did not differ as a function of test format and perceived predictive validity did not appear to serve as a partial mediator of the race/test performance relationship.

*Belief in Tests*

A number of regressions were run to test Hypotheses 7a–7d (see Table 11). Hypothesis 7a predicted a significant, positive relationship between belief in tests and test performance. In support of Hypothesis 7a, there was a significant relationship between belief in tests and test performance ($R^2 = .02$, $p < .001$). Next, Hypothesis 7b predicted a significant relationship between belief in tests and test format such that the belief in tests/multiple–choice test performance relationship would be stronger than the belief in tests/constructed response test performance relationship. To test this hypothesis, belief in tests was regressed on the test format × test performance interaction term which was not significant ($R^2 = .00$, *ns*). Consistent with this, the difference between the belief in tests/multiple–choice test performance correlation ($r = .15$, $p < .01$, 95% CI = .02 to .28) and the belief in tests/constructed response test performance correlation ($r = .19$, $p < .001$, 95% CI = .07 to .31) was not significant ($z = .44$, *ns*). Therefore, the relationship between belief in tests and test performance did not differ as a function of test format. Hypothesis 7c predicted that there would be significant subgroup differences on belief in tests. Given that ratings of belief in tests were made in reaction to having been exposed to one of the two testing conditions, the interaction between race and test format was regressed onto belief in tests. However, the results presented in Table 11 show that there was no relationship between race and belief in tests as a function of test format ($\beta = .10$, *ns*).

Table 11
*Results of Regressions for Belief in Tests (Hypotheses 7a–7d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 7a | Test Performance | Belief[A] | .15* | .02* | |
| 7b | Belief | Test Format × Test Performance | .00 | .00 | |
| 7c | Belief | Race | -.07 | | |
| | Belief | Test Format | -.12 | | |
| | Belief | Test Format × Race | .10 | .01 | |
| 7d | Test Performance | Race | .43* | **.18*** | |
| | Belief | Race | .00 | .00 | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Belief | .15* | .02* | |
| Step 2 | Test Performance | Belief | .15* | .02* | |
| | | Race | .43* | .20* | **.18*** |

*Note.* [A]Belief in tests. Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .001$.

Finally, Hypothesis 7d predicted that belief in tests would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test. The results in Table 11 indicate that the relationship between race and belief in tests was not significant, so the criteria for mediation were not met.

In summary, support was obtained for only Hypothesis 7a in that there was a significant, positive relationship between belief in tests and test performance. However, there was no support for Hypotheses 7b, 7c and 7d. As such, belief in tests did not

differ as a function of test format, there were no observed subgroup differences in belief in tests, and belief in tests did not appear to serve as a partial mediator of the race/test performance relationship.

*Stereotype Threat*

Hypotheses 8a–8d were tested by running a series of regressions (see Table 12). Hypothesis 8a predicted a significant, positive relationship between stereotype threat and test performance. In support of Hypothesis 8a, there was a significant relationship between stereotype threat and test performance ($R^2$ = .10, $p$ < .001). Next, Hypothesis 8b predicted a significant relationship between stereotype threat and test format such that the stereotype threat/multiple–choice test performance relationship would be stronger than the stereotype threat/constructed response test performance relationship. To test this hypothesis, stereotype threat was regressed on the test format × test performance interaction term which was not significant ($R^2$ = .00, *ns*). Consistent with this, the difference between the stereotype threat/multiple–choice test performance correlation ($r$ = -.27, $p$ < .001, 95% CI = -.39 to -.14) and the stereotype threat/constructed response test performance correlation ($r$ = -.41, $p$ < .001, 95% CI = -.51 to -.30) was not significant ($z$ = 1.70, *ns*). Therefore, the relationship between stereotype threat and test performance did not differ as a function of test format. Hypothesis 8c predicted that there would be significant subgroup differences on stereotype threat. Given that ratings of stereotype threat were made in reaction to having been exposed to one of the two testing conditions, the interaction between race and test format was regressed onto stereotype threat. However, the results presented in

Table 12 show that there was no relationship between race and stereotype threat as a function of test format ($\beta$ = .01, *ns*).

Table 12
*Results of Regressions for Stereotype Threat (Hypotheses 8a–8d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 8a | Test Performance | Stereotype Threat | -.32* | .10* | |
| 8b | Stereotype Threat | Test Format × Test Performance | -.03 | .00 | |
| 8c | Stereotype Threat | Race | -.08 | | |
| | Stereotype Threat | Test Format | .02 | | |
| | Stereotype Threat | Test Format × Race | .01 | .01 | |
| 8d | Test Performance | Race | .43* | **.18*** | |
| | Stereotype Threat | Race | -.08 | .01 | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Stereotype Threat | -.32* | .10* | |
| Step 2 | Test Performance | Stereotype Threat | -.29* | .10* | |
| | | Race | .41* | .27* | **.17*** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .001$.

Finally, Hypothesis 8d predicted that stereotype threat would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test. The results presented in Table 12 indicate that the criteria for mediation were not met.

In summary, support was obtained for only Hypothesis 8a in that there was a significant, positive relationship between stereotype threat and test performance. However, there was no support for Hypotheses 8b, 8c and 8d. As such, stereotype threat did not differ as a function of test format, there were no observed subgroup differences in stereotype threat, and stereotype threat did not appear to serve as a partial mediator of the race/test performance relationship.

*Self–Efficacy*

A number of regressions were run to test Hypotheses 9a–9d (see Table 13). Hypothesis 9a predicted a significant, positive relationship between self–efficacy and test performance. In support of Hypothesis 9a, there was a significant relationship between self–efficacy and test performance ($R^2$ = .08, $p$ < .001). Next, Hypothesis 9b predicted a significant relationship between self–efficacy and test format such that the self–efficacy/multiple–choice test performance relationship would be stronger than the self–efficacy/constructed response test performance relationship. To test this hypothesis, self–efficacy was regressed on the test format × test performance interaction term which was not significant ($R^2$ = .00, *ns*). Consistent with this, the difference between the self–efficacy/multiple–choice test performance correlation ($r$ = .22, $p$ < .001, 95% CI = .09 to .34) and the self–efficacy/constructed response test performance correlation ($r$ = .35, $p$ < .001, 95% CI = .23 to .46) was not significant ($z$ = 1.51, *ns*). Therefore, the relationship between self–efficacy and test performance did not differ as a function of test format. Hypothesis 9c predicted that there would be significant subgroup differences on self–efficacy. Given that ratings of self–efficacy were made in reaction to having been exposed to one of the two testing conditions, the interaction

between race and test format was regressed onto self–efficacy. The results presented in Table 13 show that there was a significant relationship between race and self–efficacy ($\beta$ = -.12, $p < .05$). Specifically, African Americans reported higher levels of self–efficacy ($M$ = 3.38, $SD$ = .78) than Whites ($M$ = 3.20, $SD$ = .79). However, the relationship between race and self–efficacy did not differ as a function of test format ($\beta$ = .07, *ns*).

Table 13
*Results of Regressions for Self–Efficacy (Hypotheses 9a–9d)*

| Hypothesis | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| 9a | Test Performance | Self–Efficacy | .28** | .08** | |
| 9b | Self–Efficacy | Test Format × Test Performance | .06 | .00 | |
| 9c | Self–Efficacy | Race | -.12* | | |
| | Self–Efficacy | Test Format | -.06 | | |
| | Self–Efficacy | Test Format × Race | .07 | .02 | |
| 9d | Test Performance | Race | .43** | **.18**** | |
| | Self–Efficacy | Race | -.12* | .01* | |
| Tests for mediation | | | | | |
| Step 1 | Test Performance | Self–Efficacy | .28** | .08** | |
| Step 2 | Test Performance | Self–Efficacy | .33** | .08** | |
| | | Race | .47** | .29** | **.21**** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .05$, **$p < .001$.

Finally, Hypothesis 9d predicted that self–efficacy would partially mediate the relationship between race and test performance such that when controlled, subgroup differences on the multiple–choice test will be reduced to the levels observed for the constructed response test. The results presented in Table 13 indicate that all the criteria for mediation were met. In addition, they show that there was a significant *increase* in the unique variance explained by race ($\Delta R^2$ = .21 vs. $R^2$ = .18) after controlling for self–efficacy. This pattern of results can only be interpreted within the context of reciprocal suppression (Conger, 1974). Reciprocal suppression occurs when two predictor variables are positively correlated with the criterion, but measure irrelevant variance in the criterion directionally opposite one another. The result is that prediction is improved with the linear combination of the two predictor variables over what would be expected from their zero–order correlations with the criterion. The conditions of reciprocal suppression are when (a) the two predictors have a negative correlation, and (b) both regression weights for the predictor/criterion relationships exceed the zero–order correlations when the variables are simultaneously entered into a regression equation. The pattern of results in these analyses met both of these conditions. For example, self–efficacy and race were negatively correlated ($r$ = -.12, $p < .05$) such that being African American was generally associated with higher self–efficacy scores. The regression weights for both the self–efficacy/test performance relationship ($\beta$ = .33, $r$ = .28, $p < .001$) and race/test performance relationship ($\beta$ = .47, $r$ = .43, $p < .001$) exceeded the zero–order correlations when both variables were entered into the regression equation. In this prediction equation, both race and self–efficacy were considered suppressor variables.

In summary, there was support for Hypotheses 9a and 9c, in that self–efficacy was related to test performance and there were significant subgroup differences on self–efficacy. However, there was no support for Hypotheses 9b and 9d.

*Criterion–Related Validity*

Hypothesis 10a posited that the multiple–choice/college GPA relationship would be the same as the constructed response/college GPA relationship. Table 4 shows that the correlation between multiple–choice test performance and college GPA ($r$ = .13, $p$ < .01, 95% CI = .00 to .26) was lower than the correlation between the constructed response test and college GPA ($r$ = .20, $p$ < .001, 95% CI = .08 to .32). However, this difference was not significant ($z$ = .77, *ns*).

Hypothesis 10b predicted that the constructed response/college GPA relationship for African Americans would be the same as the constructed response/college GPA relationship for Whites. As shown in Table 14, the correlation between constructed response test performance and college GPA for African Americans ($r$ = .29, $p$ < .001, 95% CI = .09 to .47) was higher than the correlation between constructed response test performance and college GPA for Whites ($r$ = .18, $p$ < .05, 95% CI = .02 to .33). However, this difference was not significant ($z$ = .86, *ns*). The results presented in Table 14 also indicate that the correlation between multiple–choice test performance and college GPA for African Americans ($r$ = .09, *ns*, 95% CI = -.10 to .27) was lower than the correlation between multiple–choice test performance and college GPA for Whites ($r$ = .13, *ns*, 95% CI = -.06 to .32). However, neither correlation was significant.

It was anticipated that many freshmen participants ($n$ = 132) would not have cumulative college GPAs so that much of the sample would be missing data on this

Table 14
*Correlations between Multiple–Choice and Constructed Response Test Performance and the Criterion Measures for African Americans and Whites*

| | Multiple–Choice | | Constructed Response | |
|---|---|---|---|---|
| Criteria | African American ($n$ = 115) | White ($n$ = 107) | African American ($n$ = 89) | White ($n$ = 153) |
| College GPA | .09 | .13 | .29** | .18* |
| High School GPA | .32*** | .32*** | .42*** | .31*** |
| High School Rank | .24* | .29** | .18 | .20* |

Note: *$p$ < .05, **$p$ < .01, ***$p$ < .001.

variable. As such, it was deemed necessary to collect additional criterion data such as high school GPA and high school rank in the event that a large percentage of the sample did not provide information on college GPA. However, most of the data were collected in the spring semester so the majority of freshmen had at least one semester of college grades on which to estimate their college GPA. In fact, only 27 participants (6% of the total sample) were missing data for college GPA, although 23 of the 27 were indeed freshman. Nevertheless, the regression–based approach to imputing missing data scores is accurate with less than 10% of the data missing and when the data are missing non–randomly (Roth, 1994), as was the case with the present data set. Therefore, the decision was made to impute missing scores for college GPA and use this variable as the primary criterion. However, data were gathered for high school GPA and high school rank so these variables were also used to test Hypotheses 10a and 10b.

As shown in Table 4, the correlation between the multiple–choice test performance and high school GPA ($r = .36$, $p < .001$, 95% CI = .24 to .47) was lower than the correlation between the constructed response test performance and high school GPA ($r = .46$, $p < .001$, 95% CI = .35 to .55). However, this difference was not significant ($z = 1.29$, *ns*). As shown in Table 14, the correlation between constructed response test performance and high school GPA for African Americans ($r = .42$, $p < .001$, 95% CI = .23 to .58) was higher than the correlation between constructed response test performance and high school GPA for Whites ($r = .31$, $p < .001$, 95% CI = .16 to .45). However, this difference was not significant ($z = .94$, *ns*). The results presented in Table 14 also indicate that the correlation between multiple–choice test performance and high school GPA for African Americans ($r = .32$, $p < .001$, 95% CI = .15 to .47) was the same as the correlation between multiple–choice test performance and high school GPA for Whites ($r = .32$, $p < .001$, 95% CI = .14 to .48).

Finally, the correlation between the multiple–choice test and high school rank ($r = .30$, $p < .001$, 95% CI = .18 to .41) was higher than the correlation between the constructed response test and high school rank ($r = .27$, $p < .001$, 95% CI = .15 to .38). However, this difference was not significant ($z = .35$, *ns*). As shown in Table 14, the correlation between constructed response test performance and high school rank for African Americans ($r = .18$, *ns*, 95% CI = -.02 to .37) was comparable to the correlation between constructed response test performance and college rank for Whites ($r = .20$, $p < .05$, 95% CI = .04 to .35). Therefore, using high school rank as the criterion, the overlapping confidence intervals and significance test ($z = .15$, *ns*) suggest that the criterion–related validity of the constructed response test was the same for both African

Americans and Whites. The results presented in Table 14 also indicate that the correlation between multiple–choice test performance and high school rank for Whites ($r$ = .29, $p$ < .01, 95% CI = .11 to .45) was higher than the correlation between multiple–choice test performance and high school rank for African Americans ($r$ = .24, $p$ < .05, 95% CI = .06 to .40). However, this difference was also not significant ($z$ = .40, *ns*).

*Supplementary Analyses*

The model presented in Figure 1 posits that each of the six test perception variables (i.e., face validity, fairness, perceived predictive validity, belief in tests, stereotype threat, and self–efficacy) is related to all of the other test perception variables. In addition, the model indicates that face validity, fairness, perceived predictive validity, and belief in tests mediate the relationship between race and motivation with stereotype threat and self–efficacy mediating the relationship between race and motivation and anxiety. It was also posited that test–taking motivation and anxiety are directly related to test performance. To examine these relationships that are implied in the model presented in Figure 1, supplementary analyses were conducted for each of the six test perception variables.

*Face Validity*

An examination of the correlations in Table 4 shows that face validity was significantly related to fairness, perceived predictive validity, and belief in tests in both test format conditions, but related to self–efficacy in only the constructed response test condition. Face validity was not related to stereotype threat in either test format condition. The magnitude of the correlations ranged from $r$ = -.04 to .43 (absolute

mean $r$ = .22, $SD$ = .16) in the multiple–choice test condition and ranged from $r$ = -.09 to .54 (absolute mean $r$ = .34, $SD$ = .20) in the constructed response test condition. In general, the relationships between face validity and the other perception variables were stronger in the constructed response than the multiple–choice test condition (with the exception of stereotype threat).

The model in Figure 1 posits that face validity mediates the relationship between race and test performance, only through its relationship with motivation which is directly related to test performance. To test these relationships, two sets of regressions were computed in accordance with standards outlined by Baron and Kenny (1986). The first set of regressions tested the extent to which face validity mediated the relationship between race and motivation. As shown in Table 15, all three criteria for mediation were met, but there was no decrease in the unique variance explained by race ($\Delta R^2$ = .03 vs. $R^2$ = .03) after controlling for face validity. Therefore, face validity did not mediate the relationship between race and motivation. The second set of regressions tested the extent to which motivation mediated the relationship between face validity and test performance. As shown in Table 15, the criteria for mediation were not met and hence, there was no evidence that motivation mediated the relationship between face validity and test performance.

Table 15
*Mediation Tests for the Relationship among Face Validity, Motivation, Anxiety, and Test Performance from Figure 1*

| Relationship | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| Race Face Validity Motivation | Motivation | Race | -.19*** | **.03**** | |
| | Face Validity | Race | -.10* | .01* | |
| Step 1 | Motivation | Face Validity | .16*** | .02*** | |
| Step 2 | Motivation | Face Validity | .14** | .02*** | |
| | | Race | -.17*** | .05*** | **.03**** |
| Face Validity Motivation Test Performance | Test Performance | Face Validity | .00 | **.00** | |
| | Motivation | Face Validity | .16*** | .02*** | |
| Step 1 | Test Performance | Motivation | .05 | .00 | |
| Step 2 | Test Performance | Motivation | .05 | .00 | |
| | | Face Validity | -.01 | .00 | **.00** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .05$, **$p < .01$, ***$p < .001$.

*Fairness*

An examination of the correlations in Table 4 shows that fairness was significantly related to all other test perception variables. The magnitude of the correlations ranged from $r = .21$ to .46 (absolute mean $r = .37$, $SD = .10$) in the multiple–choice test condition and ranged from $r = -.33$ to .56 (absolute mean $r = .42$, $SD = .10$) in the constructed response test condition. In general, the relationships between fairness and the other perception variables were stronger in the constructed response than the multiple–choice test condition (with the exception of self–efficacy).

The model in Figure 1 posits that fairness mediates the relationship between race and test performance, only through its relationship with motivation which is directly related to test performance. To test these relationships, two sets of regressions were computed in accordance with standards outlined by Baron and Kenny (1986). The first set of regressions tested the extent to which fairness mediated the relationship between race and motivation. As shown in Table 16, all three criteria for mediation were met, but there was an increase in the unique variance explained by race ($\Delta R^2$ = .05 vs. $R^2$ = .03) after controlling for fairness which indicates that fairness was suppressing the relationship between race and motivation. The second set of regressions tested the extent to which motivation mediated the relationship between fairness and test performance. As shown in Table 16, the criteria for mediation were not met and hence, there was no evidence that motivation mediated the relationship between fairness and test performance.

Table 16
*Mediation Tests for the Relationship among Fairness, Motivation, Anxiety, and Test Performance from Figure 1*

| Relationship | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| Race Fairness Motivation | Motivation | Race | -.19** | **.03**** | |
| | Fairness | Race | .12* | .01* | |
| Step 1 | Motivation | Fairness | .24** | .06** | |
| Step 2 | Motivation | Fairness | .27** | .06** | |
| | | Race | -.22** | .11** | **.05**** |
| Fairness Motivation Test Performance | Test Performance | Fairness | .21** | **.04**** | |
| | Motivation | Fairness | .24** | .06** | |
| Step 1 | Test Performance | Motivation | .05 | .00 | |
| Step 2 | Test Performance | Motivation | .00 | .00 | |
| | | Fairness | .21** | .04** | **.04**** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .01$, **$p < .001$.

*Perceived Predictive Validity*

An examination of the correlations in Table 4 shows that perceived predictive validity was significantly related to all other test perception variables. The magnitude of the correlations ranged from $r = .19$ to .57 (absolute mean $r = .38$, $SD = .16$) in the multiple–choice test condition and ranged from $r = -.18$ to .62 (absolute mean $r = .43$, $SD = .18$) in the constructed response test condition. In general, the relationships between perceived predictive validity and the other perception variables were stronger in the constructed response than the multiple–choice test condition (with the exception of stereotype threat).

The model in Figure 1 posits that perceived predictive validity mediates the relationship between race and test performance, only through its relationship with motivation which is directly related to test performance. To test these relationships, two sets of regressions were computed in accordance with standards outlined by Baron and Kenny (1986). The first set of regressions tested the extent to which perceived predictive validity mediated the relationship between race and motivation. The second set of regressions tested the extent to which motivation mediated the relationship between perceived predictive validity and test performance. As shown in Table 17, the criteria for mediation for both sets of analyses were not met and hence, there was no evidence for the relationships regarding perceived predictive validity as implied in Figure 1.

Table 17
*Mediation Tests for the Relationship among Perceived Predictive Validity, Motivation, Anxiety, and Test Performance from Figure 1*

| Relationship | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| Race Predictive Motivation | Motivation | Race | -.19** | **.04**** | |
| | Predictive | Race | .00 | .00 | |
| Step 1 | Motivation | Predictive | .15** | .02** | |
| Step 2 | Motivation | Predictive | .15** | .02** | |
| | | Race | -.19*** | .06*** | **.04***** |
| Predictive Motivation Test Performance | Test Performance | Predictive | .12** | **.02**** | |
| | Motivation | Predictive | .15** | .02** | |
| Step 1 | Test Performance | Motivation | .05 | .00 | |
| Step 2 | Test Performance | Motivation | .04 | .00 | |
| | | Predictive | .12* | .02* | **.02*** |

*Note.* [A]Perceived predictive validity. Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .05$, **$p < .01$, ***$p < .001$.

*Belief in Tests*

An examination of the correlations in Table 4 shows that belief in tests was significantly related to all other test perception variables. The magnitude of the correlations ranged from $r = .34$ to .57 (absolute mean $r = .41$, $SD = .10$) in the multiple–choice test condition and ranged from $r = -.36$ to .62 (absolute mean $r = .49$, $SD = .12$) in the constructed response test condition. In general, the relationships between belief in tests and the other perception variables were stronger in the constructed response than the multiple–choice test condition (with the exception of stereotype threat).

The model in Figure 1 posits that belief in tests mediates the relationship between race and test performance, only through its relationship with motivation which is directly related to test performance. To test these relationships, two sets of regressions were computed in accordance with standards outlined by Baron and Kenny (1986). The first set of regressions tested the extent to which belief in tests mediated the relationship between race and motivation. The second set of regressions tested the extent to which motivation mediated the relationship between belief in tests and test performance. As shown in Table 18, the criteria for mediation for both sets of analyses were not met and hence, there was no evidence for the relationships regarding belief in tests as implied in Figure 1.

Table 18
*Mediation Tests for the Relationship among Belief in Tests, Motivation, Anxiety, and Test Performance from Figure 1*

| Relationship | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| Race Belief Motivation | Motivation | Race | -.19** | **.03**** | |
| | Belief | Race | .00 | .00 | |
| Step 1 | Motivation | Belief | .22** | .05** | |
| Step 2 | Motivation | Belief | .22** | .05** | |
| | | Race | -.19** | .08** | **.03**** |
| Belief Motivation Test Performance | Test Performance | Belief | .15** | **.02**** | |
| | Motivation | Belief | .22** | .05** | |
| Step 1 | Test Performance | Motivation | .05 | .00 | |
| Step 2 | Test Performance | Motivation | .02 | .00 | |
| | | Belief | .15* | .02* | **.02*** |

*Note.* [A]Belief in tests. Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .01$, **$p < .001$.

*Stereotype Threat*

An examination of the correlations in Table 4 shows that stereotype threat was significantly related to all of the perception variables except for face validity. The magnitude of the correlations ranged from $r = -.09$ to -.58 (absolute mean $r = .32$, $SD = .18$) in the multiple–choice test condition and ranged from $r = -.09$ to -.60 (absolute mean $r = .31$, $SD = .20$) in the constructed response test condition. The relationships between stereotype threat and the other perception variables were generally equal in magnitude across the constructed response and the multiple–choice test conditions.

The model in Figure 1 posits that stereotype threat mediates the relationship between race and test performance, only through its relationship with anxiety and motivation which are directly related to test performance. To test these relationships as posited in Figure 1, four sets of regressions were computed in accordance with standards outlined by Baron and Kenny (1986). The first two sets of regressions tested the extent to which stereotype threat mediated the relationship between race and anxiety and motivation. However, as shown in Table 19, the criteria for mediation for both sets of analyses were not met and hence there was no evidence for these relationships. It is important to note that the pattern of results for the relationships among race, stereotype threat and anxiety indicate that stereotype threat may be suppressing the relationship between race and anxiety (i.e., $\Delta R^2 = .01$ vs. $R^2 = .00$) . However, the criteria for reciprocal suppression were not met.

The last two sets of regressions tested the extent to which anxiety and motivation mediated the relationship between stereotype threat and test performance. As shown in Table 19, there was evidence to support the position that anxiety partially mediated the relationship between stereotype threat and test performance. For example, all three criteria for mediation were met, there was a significant decrease in the unique variance explained by stereotype threat ($\Delta R^2 = .02$ vs. $R^2 = .10$) after controlling for anxiety, and Sobel's test for the indirect effect of stereotype threat and test performance through anxiety was significant (Sobel = 2.83, $p < .01$). There was no evidence that motivation mediated the relationship between stereotype threat and test performance. Although the criteria for mediation were met, there was no decrease in the unique

Table 19
*Mediation Tests for the Relationship among Stereotype Threat, Motivation, Anxiety, and Test Performance from Figure 1*

| Relationship | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| Race Stereo Anxiety | Anxiety | Race | .06 | **00** | |
| | Stereotype Threat | Race | -.08 | .01 | |
| Step 1 | Anxiety | Stereotype Threat | .66** | .44** | |
| Step 2 | Anxiety | Stereotype Threat | .67** | .44** | |
| | | Race | .11* | .45* | **.01*** |
| Race Stereo Motivation | Motivation | Race | -.19* | **.04**** | |
| | Stereotype Threat | Race | -.08 | .01 | |
| Step 1 | Motivation | Stereotype Threat | -.21** | .04** | |
| Step 2 | Motivation | Stereotype Threat | -.23** | .04** | |
| | | Race | -.20** | .09** | **.04**** |
| Stereo Anxiety Performance | Test Performance | Stereotype Threat | -.32** | **.10**** | |
| | Anxiety | Stereotype Threat | .66** | .44** | |
| Step 1 | Test Performance | Anxiety | -.31** | .10** | |
| Step 2 | Test Performance | Anxiety | -.18* | .10** | |
| | | Stereotype Threat | -.21** | .12** | **.02**** |
| Stereo Motivation Performance | Test Performance | Stereotype Threat | -.32** | **.10*** | |
| | Motivation | Stereotype Threat | -.21** | .04** | |
| Step 1 | Test Performance | Motivation | .05 | .00 | |
| Step 2 | Test Performance | Motivation | -.01 | .00 | |
| | | Stereotype Threat | -.33** | .10** | **.10**** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .01$, **$p < .001$.

variance explained by stereotype threat ($\Delta R^2$ = .10 vs. $R^2$ = .10) after controlling for motivation.

In summary, a test of the relationships outlined in Figure 1 regarding stereotype threat indicated that stereotype threat was related to all the other test perceptions and anxiety partially mediated the relationship between stereotype threat and test performance. However, the results did not support the other relationships described in Figure 1 regarding stereotype threat.

*Self–Efficacy*

An examination of the correlations in Table 4 shows that self–efficacy was significantly related to all of the other perception variables except for face validity. The magnitude of the correlations ranged from $r$ = -.04 to -.58 (absolute mean $r$ = .31, $SD$ = .21) in the multiple–choice test condition and ranged from $r$ = .20 to -.60 (absolute mean $r$ = .37, $SD$ = .15) in the constructed response test condition. In general, the relationships between self–efficacy and the other perception variables were stronger in the constructed response than the multiple–choice test condition (with the exception of fairness).

The model in Figure 1 posits that self–efficacy mediates the relationship between race and test performance, only through its relationship with anxiety and motivation which are directly related to test performance. To test these relationships, four sets of regressions were computed in accordance with standards outlined by Baron and Kenny (1986). The first two sets of regressions tested the extent to which self–efficacy mediated the relationship between race and anxiety and motivation. As shown in Table 20, there was no evidence that

self–efficacy mediated the relationship between race and anxiety because the criteria for mediation were not met. There was some evidence to support the position that self–efficacy partially mediated the relationship between race and motivation. For example, all three criteria for mediation were met, there was a significant decrease in the unique variance explained by race ($\Delta R^2$ = .02 vs. $R^2$ = .03) after controlling for self–efficacy, and Sobel's test for the indirect effect of race and motivation through self–efficacy was significant (Sobel = 3.92, $p < .001$).

The last two sets of regressions tested the extent to which anxiety and motivation mediated the relationship between self–efficacy and test performance. As shown in Table 20, there was evidence to support the position that anxiety fully mediated the relationship between self–efficacy and test performance. For example, all three criteria for mediation were met, there was a significant decrease in the unique variance explained by self–efficacy ($\Delta R^2$ = .00 vs. $R^2$ = .08) after controlling for anxiety, and Sobel's test for the indirect effect of self–efficacy and test performance through anxiety was significant (Sobel = 3.01, $p < .01$). Although the pattern of results for the last set of regressions indicates that motivation may be suppressing the relationship between self–efficacy and test performance ($\Delta R^2$ = .09 vs. $R^2$ = .08) , the criteria for reciprocal suppression were not met.

In summary, a test of the relationships outlined in Figure 1 regarding self–efficacy indicated that self–efficacy was related to all the other test perceptions, self–efficacy partially mediated the relationship between race and motivation, and anxiety fully mediated the relationship between self–efficacy and test performance.

Table 20
*Mediation Tests for the Relationship among Self–Efficacy, Motivation, Anxiety, and Test Performance from Figure 1*

| Relationship | Dependent Variable | Independent Variable | β | $R^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| Race Self–Efficacy Anxiety | Anxiety | Race | .06 | **.00** | |
| | Self–efficacy | Race | -.12* | .01* | |
| Step 1 | Anxiety | Self–efficacy | -.76** | .58** | |
| Step 2 | Anxiety | Self–efficacy | -.76** | .58** | |
| | | Race | -.03 | 58** | **.00** |
| Race Self–Efficacy Motivation | Motivation | Race | -.19** | **.03**** | |
| | Self–efficacy | Race | -.12* | .01* | |
| Step 1 | Motivation | Self–efficacy | .44** | .20** | |
| Step 2 | Motivation | Self–efficacy | .43** | .19** | |
| | | Race | -.14** | .21** | **.02**** |
| Self–Efficacy Anxiety Test Performance | Test Performance | Self–efficacy | .28** | **.08**** | |
| | Anxiety | Self–efficacy | -.76** | .58** | |
| Step 1 | Test Performance | Anxiety | -.31** | .10** | |
| Step 2 | Test Performance | Anxiety | -.23** | .10** | |
| | | Self–efficacy | .10 | .10** | **.00** |
| Self–Efficacy Motivation Test Performance | Test Performance | Self–efficacy | .28** | **.08**** | |
| | Motivation | Self–efficacy | .44** | .20** | |
| Step 1 | Test Performance | Motivation | .05 | .00 | |
| Step 2 | Test Performance | Motivation | -.09 | .00 | |
| | | Self–efficacy | .32** | .09** | **.09**** |

*Note.* Numbers in bold represent the comparison between $\Delta R^2$ and $R^2$ for the test for mediation. *$p < .05$, **$p < .001$.

# DISCUSSION AND CONCLUSIONS

The objectives of the present study were to (a) replicate the results of Arthur et al. (2002) with a larger sample size, (b) empirically examine factors that may explain a reduction in race–based subgroup differences observed on a constructed response compared to a multiple–choice test, and (c) assess the criterion–related validity of the constructed response test.

In general, the *pattern* of results supported the hypotheses in the predicted direction. For example, although there was a reduction in subgroup differences in performance on the constructed response compared to the multiple–choice test, the difference was not statistically significant. However, analyses by specific test content yielded a significant reduction in subgroup differences on the science reasoning section. In addition, all of the hypothesized study variables, with the exception of face validity, were significantly related to test performance. Significant subgroup differences were also obtained for all study variables except belief in tests and stereotype threat. Further analyses of the perceptions variables indicated that ratings of test perceptions for Whites and African Americans differed across the two test formats in the predicted direction, although significant effects were obtained only for fairness and perceived predictive validity. The results also indicate that reading ability, test–taking skills, and perceived fairness partially mediated the relationship between race and test performance. Finally, the criterion–related validity for the constructed response test was stronger than that for the multiple–choice test.

Efforts to reduce subgroup performance differences in high–stakes testing are predicated on the assumption that reductions in subgroup differences will ultimately

lead to reductions in adverse impact for members of protected classes. A test or assessment tool displays adverse impact if there are differential outcomes associated with the use of the test (e.g., for selection, promotion) as a function of a protected class status variable (e.g., race, sex). Adverse impact is typically operationalized in terms of the 80% or 4/5th rule. That is, a selection rate for any race, sex, or ethnic group that is less than 4/5th or 80% of the rate for the group with the highest rate constitutes adverse impact (EEOC, 1978). The degree to which subgroup differences leads to adverse impact is largely a function of the cutoff score used to determine which test takers "pass" or "fail" the test. Thus, in the presence of subgroup differences, how the cutoff score is established plays a critical role in determining the presence and level of adverse impact.

Although 70% is a widely used cutoff score in municipal testing, for exploratory purposes, cutoff scores at the mean and one standard deviation below the mean were also examined to assess the levels of adverse impact on the multiple–choice and constructed response tests. In addition, two separate regression–based cutoff scores for each test were established by regressing college GPA and high school GPA onto multiple–choice and constructed response test performance. The cutoff score for each test was estimated for college GPA and high school GPA of 2.0 which is considered the minimum GPA for passing. These two criteria were chosen because college GPA was the primary criterion and among all three criterion measures, high school GPA had the strongest relationship with test performance. The actual raw score cutoff scores for each test were as follows: 70%, multiple–choice = 14, constructed response = 14; mean, multiple–choice = 13.4, constructed response = 11.1; one standard deviation below the

mean, multiple–choice = 9.2, constructed response = 6.9; regression–based cutoff score using college GPA, multiple–choice = 13.5, constructed response = 10.1; and regression–based cutoff score using high school GPA, multiple–choice = 10.1, constructed response = 3.52.

As seen in Figure 7, subgroup differences on both tests resulted in adverse impact against African Americans using four of the five cutoff scores examined. However, the regression–based cutoff scores using high school GPA as the criterion yielded no adverse impact against African Americans. Furthermore, with the exception of the cutoff scores at the mean, levels of adverse impact were lower for the constructed response test Therefore, the reductions in subgroup differences observed for the constructed response test translated to less adverse impact than for the multiple–choice test using four of the five cutoff scores.

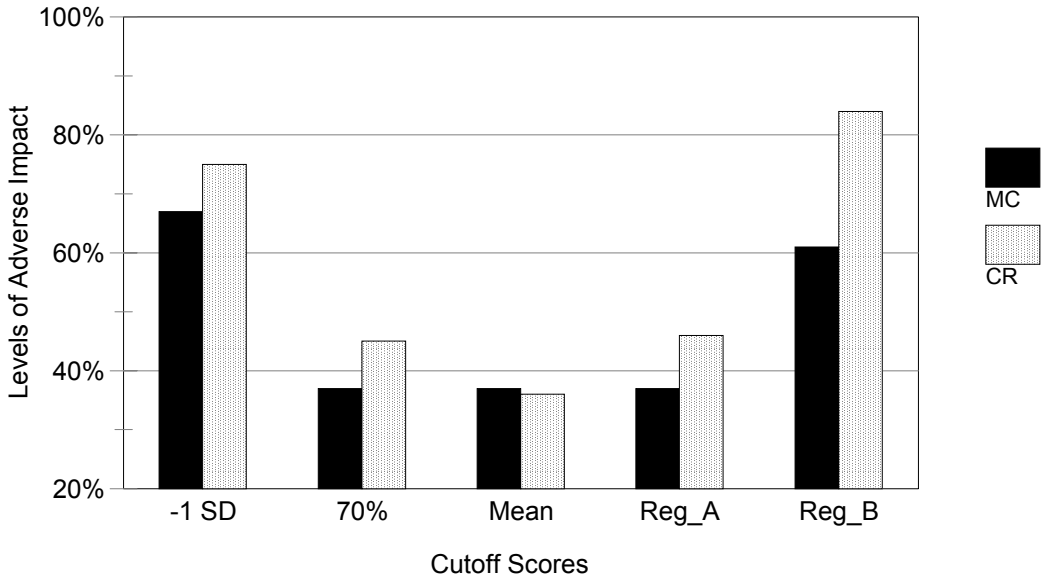


*Figure 7*. Levels of adverse impact for the multiple–choice and constructed response tests. Reg_A = regression–based cutoff score using college GPA and Reg_B = regression–based cutoff score using high school GPA.

This study contributes to the extant literature concerning the use of alternative test formats in efforts to reduce subgroup differences on cognitively loaded paper–and–pencil tests of knowledge, skills, and abilities. Although not statistically significant, the pattern of results suggest reductions in subgroup differences were obtained on the constructed response format compared to the multiple–choice format. Furthermore, the reduction in subgroup differences translated into lower levels of adverse impact on the constructed response test for four of five cutoff scores examined. The results also suggest that this reduction in subgroup differences on the constructed response test may be partially explained by differences in the reading load, susceptibility to testwiseness, and some test perceptions associated with the two test

formats. Although only two of the test perceptions variables (i.e., fairness and perceived predictive validity) yielded significant subgroup differences across test format, the means for all other test perception variables were in the predicted direction. For example, the trend in the data suggest that African Americans report higher levels of self–efficacy and a belief in tests than Whites in the constructed response test condition and view the constructed response test as being more face valid, fairer, and having more predictive validity than Whites. Therefore, in general, African Americans reported more favorable test perceptions for the constructed response test than the multiple–choice test. Prior investigations of the viability of alternative test formats have compared paper–and–pencil and high fidelity test formats such as performance tests, job simulations, or multi–media presentations. However, these alternative test formats have lower economic utility than paper–and–pencil tests because they require more resources to develop, administer, and score. Therefore, another advantage of the constructed response test format presented in the present study is that it preserves the advantages of paper–and–pencil tests. Finally, there was no loss in the criterion–related validity with the use of this alternative format. In fact, the criterion–related validity for the constructed response test in the present study was higher than the criterion–related validity for the multiple–choice test. Thus, the constructed response format may be a viable alternative to the traditional paper–and–pencil multiple–choice format in high–stakes testing in solving the organizational dilemma of using the most valid predictors of job performance and simultaneously reducing subgroup differences and subsequent adverse impact on tests of knowledge, skill, ability, and achievement.

The present research study addressed a common weakness in the extant literature by drawing a distinction between constructs and methods. Comparisons of test methods should hold the construct constant to obtain meaningful results. To maximize construct equivalency on the two tests used in the present study, the items on each test were stem–equivalent. Holding the construct constant and varying only the method used to measure the specified construct allowed for the systematic examination of the impact of test format on subgroup differences.

Although the items in each test were stem–equivalent, it could be argued that the two test formats measured substantively different constructs in addition to the intended constructs (i.e., mathematics and science reasoning). Indeed, the primary impetus for the proposed study is that multiple–choice tests may measure irrelevant constructs such as reading ability, test–taking skills, or test–perceptions (Arthur et al., 2002; Chan & Schmitt, 1997; Schmitt et al., 1996) in addition to the relevant construct. Therefore, it was not expected that there was 100% overlap between the two tests in regards to construct measurement. Any variance in test performance attributed to irrelevant constructs such as reading ability, test–taking skills, and test perceptions may be different across the two test formats. In addition, some prior evidence suggests that the same construct may not be measured similarly by two different test formats (Traub, 1993). Consistent with this, subsequent analyses by test construct in the present study indicated that subgroup differences were significantly reduced on the constructed response test for the science reasoning section, but not the math section. However, it is important to note that this effect was obtained with a decrease in test performance on the constructed response test.

Based on the present results, it is plausible that math tests are less influenced by test format because the measurement of extraneous variables is minimized due to the nature of responses to math questions that require computations, regardless of the response format. In contrast, constructs such as science reasoning may introduce the measurement of extraneous variables (e.g., reading ability and test–taking skills) that contaminate predictor scores and are therefore more likely to be influenced by changing the test format. Therefore, the constructed response test format may yield greater reductions in subgroup differences for constructs that are structurally similar to science reasoning than math.

Another contribution of the present study was the examination of factors—reading ability, test–taking skills, and test perceptions—that might explain why changes in test format may reduce subgroup differences. For example, based on the extant literature (e.g., Chan & Schmitt, 1997; Sacco et al., 2000), it was expected that reading ability would be one of the strongest explanatory variables for the reduction in subgroup differences. Next, several studies have shown that perceptions of and reactions to specified tests are related to test performance with some research revealing the existence of subgroup differences on various test perceptions (Ryan, 2001). Third, a review of the literature revealed weak and sometimes inconclusive results on the relationship among race, test–taking skills, and test performance. Thus, it was difficult, based on prior research, to anticipate the role of test–taking skills in the relationship between race and test performance.

Because of differences in the reading load of the two test formats, it was predicted that reading ability would partially mediate the race/test performance

relationship. All but one of the hypotheses were supported regarding reading ability. Specifically, there was a strong relationship between reading ability and test performance (Hypothesis 2a) and there were subgroup differences in reading ability (Hypothesis 2c). However, the reading ability/multiple–choice test performance relationship did not differ from the reading ability/constructed response test performance relationship (Hypothesis 2b). Finally, evidence for partial mediation of reading ability was obtained, but the strength of the effect was not sufficient to significantly reduce levels of subgroup differences in multiple–choice test performance to the levels observed on the constructed response test after statistically controlling for reading ability (Hypothesis 2d).

Prior research (Benson et al., 1986; Diamond et al., 1976; Rogers & Yang, 1996) suggests that Whites and African Americans do not differ in test–taking skills. In addition, the influence of training for test–taking skills has exhibited only a weak relationship with test performance (e.g., Bangert–Drowns et al., 1983; Dolly & Williams, 1986). Nevertheless, this variable was examined in the present study because the two test formats were expected to differ in their susceptibility to testwiseness cues. The results for test–taking skills were similar to those for reading ability. That is, support was obtained for a relationship between test–taking skills and test performance (Hypothesis 3a) and there were subgroup differences in test–taking skills (Hypothesis 3c). However, the test–taking skills/multiple–choice test performance relationship did not differ from the test–taking skills/constructed response test performance relationship (Hypothesis 3b). Finally, evidence for partial mediation of test–taking skills was obtained, but the strength of the effect was not sufficient to significantly reduce the

levels of subgroup differences in multiple–choice test performance to the levels observed on the constructed response test after statistically controlling for test–taking skills (Hypothesis 3d).

Sackett et al. (2001) suggested that minimizing race–based performance differences through altering test perceptions is not likely to have a large effect on test scores. Nevertheless, perceptions were investigated in the present study since prior research (Chan et al., 1998; Ryan et al., 2000; Schmitt & Mills, 2001) suggests that test perceptions such as face validity, fairness, perceived predictive validity, and self–efficacy may be related to test performance. In addition, it was hypothesized that test perceptions would differ across the two test formats. The results indicated that all test perceptions, with the exception of face validity, were related to test performance. Subgroup differences were also obtained for fairness and perceived predictive validity. Specifically, the results indicated that African Americans perceived the constructed response test to be fairer and also have higher predictive validity than Whites. In contrast, Whites perceived the multiple–choice test to be fairer and have higher predictive validity than African Americans. Only perceived fairness appeared to partially mediate the relationship between race and test performance, but the strength for mediation was not sufficient to significantly reduce levels of subgroup differences in multiple–choice test performance to the levels observed on the constructed response test after statistically controlling for fairness.

The conceptual model presented in Figure 1 implies that the test perception variables are interrelated, that each test perception variable mediates the relationship between race and test–taking motivation and/or anxiety, and that test–taking motivation

and anxiety mediate the relationship between test perceptions and performance. Although no hypotheses were specified for these relationships, they were evaluated to test the entire model presented in Figure 1. With a few exceptions, the intercorrelations among the perception variables were significant. Tests for mediation among perceptions, test–taking motivation, test–taking anxiety, and test performance indicated that only self–efficacy mediated the relationship between race and test–taking motivation. Furthermore, the results suggested that anxiety mediated the relationship between both stereotype threat and self efficacy and test performance. No other significant relationships were obtained.

Another contribution to the extant literature is that criterion–related validity evidence was obtained for both test formats. Previous examinations of subgroup differences on alternative predictors of job performance have typically failed to provide any criterion–related validity evidence. Although a reliance on content–related validity evidence may not in and of itself be deficient, the additional demonstration of criterion–related validity further bolsters the efficacy and utility of the constructed response test. Consistent with the hypothesized effects, the criterion–related validity of the constructed response and multiple–choice tests were similar. Furthermore, the criterion–related validity for the constructed response test was similar for African Americans and Whites. This is consistent with meta–analytic findings that validity differences by race seldom occurs (Hunter, Schmidt, & Hunter, 1979; Schmidt, Pearlman, & Hunter, 1980). Consequently, there was no loss of criterion–related validity using a constructed response format as an alternative to multiple–choice tests of knowledge, skills, and abilities.

*Limitations*

The primary limitation of the present study is that reductions in subgroup differences were obtained with a decrease in test performance on the constructed response compared to the multiple–choice test. Although using the constructed response test preserved the advantages of paper–and–pencil tests and reduced subgroup differences, the cost was a decrement in performance for all test takers. This performance difference may be interpreted within the context of the well documented differences in performance between tests of recognition and recall (Anderson, 1999). That is, all things being equal, performance on recognition tests (e.g., multiple–choice) is generally better than that on recall tests (e.g., constructed response).

It is recognized that compared to multiple–choice items, the scoring of write–in items similar to those used in the constructed response test is more labor intensive and obviously introduces an element of subjectivity in the scoring process. These are not necessarily insurmountable problems, as highlighted by the extensive use of employment interviews that are by definition, very subjective and labor intensive in both administration and scoring. Furthermore, compared to the cost of developing, administering, and scoring a performance test, the relative cost of scoring a write–in test is comparably small. Finally, although constructed, the relatively constrained responses (compared to an essay) used in the specific constructed response test format presented here made it possible to readily standardize the scoring and accomplish it in an efficient and psychometrically sound manner; it took on average 55 seconds to score a test, and the preconsensus interscorer agreement was high (96% at Time 1 and 99% at Time 2).

One possible reason why subgroup differences on the constructed response test were not significantly reduced compared to the multiple–choice test is that the reading load of each test was comparable. If reading ability is an explanation for reductions in subgroup differences on tests with lower reading demands, then ideally, reading loads across the two test formats should be maximally different (e.g., performance tests vs. paper–and–pencil tests). The constructed response test in the present study contained 27% fewer words than the multiple–choice test, and this reduction in reading load may not have been sufficient to obtain significant reductions in subgroup differences. For example, although there were 24% fewer words on the math section of the constructed response than the multiple–choice test, all alternatives (and hence all of the missing "words") for the 10 items in the multiple–choice test were numbers. So, reading demands would have been higher if the alternatives were text– instead of number–based. In addition, there were 27% fewer words on the science reasoning section of the constructed response than the multiple–choice test. However, on the science reasoning section, background materials associated with the items were presented on both tests which still presented a heavy reading load on the constructed response test. This could explain why reductions in observed subgroup differences on the constructed response test in the present study were not as small as those obtained by Arthur et al. (2002) in the within–subjects design ($d$ = .12) on their constructed response test which contained no background materials (the constructed response test used in their study contained 60% fewer words than the multiple–choice test). Nevertheless, the minimal reduction in reading load on the present constructed response test was

sufficient to obtain a reduction in subgroup differences (although it did not reach significance).

It is important to note that reading ability is considered an extraneous variable only in situations in which reading ability is not a job requirement. That is, to the extent that reading ability is not required for the job, the measurement of reading ability in conjunction with job–relevant constructs reduces the construct validity of a selection or promotional test. However, in the present research design, reading ability is related to the criterion (i.e., college GPA). Therefore, the reductions in subgroup differences on the constructed response test explained by reading ability in the present academic setting are practically significant. The impact of using a constructed response test should only be stronger in selection or promotional settings where reading ability is not required.

Another limitation of the present study was the weak psychometric properties of some of the measures. Specifically, relatively low score reliabilities were obtained for the measures of test–taking skills (.57), fairness (.64), and stereotype threat (.47). Although the measure of test–taking skills used in the present study (Gibb, 1964) is generally considered the best measure of test–taking skills (Harmon et al., 1996; Miller et al., 1990), the internal consistency reported in the test manual for the full 70–item measure is only .72. Therefore, the reduction in items from 70 to 20 in the present study may have substantially altered the reliability and validity of the original test. This conclusion is bolstered by the fact that contrary to expectations, there was no relationship in the present study between test–taking skills and experience with taking multiple–choice or constructed response tests. A Spearman–Brown correction was

applied to the reliability coefficient obtained in the present study (.57 for the 20–item measure) for the measure of test–taking skills to estimate the reliability for the full 70–item measure (.82) in the present sample.

Another potential limitation of the present study is the generalizability of the findings. The present study was conducted in a laboratory setting in which there were no real consequences associated with the test scores. Although participants reported exhibiting "quite a bit of effort" and the top 20 scorers were awarded $30 to increase motivation, participants may have exhibited less effort on the test than in real world organizational settings. As such, this is a potential threat to the ecological validity of the study. It is expected that reductions in subgroup differences on constructed response tests will be larger in high–stakes testing situations for which this intervention strategy is designed.

The decision to use items from the ACT for the multiple–choice and constructed response test in the present study was based on several factors. First, this is an operational test that is used for selection into universities and colleges so using this test emulates the selection models applied by organizations in the measurement of knowledge, skill, ability, and achievement to predict future job performance. Second, use of the ACT allowed for the collection of real criterion data (i.e., college GPA) because it is a selection tool employed by universities for selection. However, a major disadvantage of the ACT in the present study is the presence of heavy reading demands which are relevant to the prediction of college performance. As a result, the reading load on the constructed response test was minimally reduced (i.e., by 27%). As noted previously, this minimal reduction in reading load may have resulted in the weak effect

sizes obtained for the relationships tested. Future comparisons of the constructed response and multiple–choice test formats in lab–based settings should use tests that more closely resemble tests of knowledge, skill, ability, and achievement used in operational job settings. These knowledge–, skill–, ability–, or achievement–based tests, especially for jobs characterized by less reading demands, typically have lower reading loads. As a result, differences in reading demands, and subsequently, subgroup differences between multiple–choice and constructed response formats may be larger. Even more desirable would be field studies which allow for the collection of research data to test explanations for reductions in subgroup differences.

Another suggestion for future research is to explore other measures of test–taking skills that are shorter than Gibb's Experimental Test of Testwiseness (1964) and have better psychometric properties than the version used in the present study. For example, Bruch (1981) asked test takers to describe the test–taking strategies they used on a specific test. These statements were then coded by experimenters into categories and given a score to represent the total number of strategies used and the relative effectiveness of the strategies. In another study, Ellis and Ryan (1999) had test takers endorse items on a checklist describing various effective and ineffective test–taking strategies participants used on the test.

In summary, the *pattern* of results supported the hypotheses in the predicted direction for most variables. For example, reductions in subgroup differences were obtained on the constructed response format compared to the multiple–choice format. Furthermore, the reductions in subgroup differences translated into lower levels of adverse impact on the constructed response test for four of five cutoff scores examined.

The results also suggested that this reduction in subgroup differences on the constructed response test may be partially explained by differences in the reading load, susceptibility to testwiseness, and some test perceptions associated with the two test formats. In addition, the results suggested that African Americans had more favorable perceptions of the constructed response than the multiple–choice test. Finally, there was no loss in the criterion–related validity with the use of this alternative format. In fact, the criterion–related validity of the constructed response test in the present study was higher than the criterion–related validity of the multiple–choice test. The constructed response format presented also retains many of the advantages of paper–and–pencil multiple–choice tests such as low cost of development, administration, and scoring, relatively objective scoring, and comparable criterion–related validity. However, it has the potential, added advantage of reducing subgroup differences in test performance via reading ability, test–taking skills, and test perceptions. Therefore, it may be a more desirable alternative to the traditional multiple–choice format in high–stakes testing in efforts to solve the organizational dilemma.

# REFERENCES

Ambady, N., Shih, M., Kim, A., & Pittinsky, T. L. (2001). Stereotype susceptibility in children: Effects of identity activation on quantitative performance. *Psychological Science, 5*, 385–390.

American College Testing (2002). *ACT assessment homepage*. Retrieved July 25th, 2002, from http://www.act.org/aap/index.html

Anderson, J. R. (1999). *Cognitive psychology and its implications.* (5th ed.) New York: Worth Publishing.

Arthur, W. Jr., Bell, S. T., & Edwards, B. D. (2003). *An empirical comparison of the criterion–related validities of additive and referent–shift operationalizations of team efficacy.* [Manuscript submitted for publication].

Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement, 54*, 394–403.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. (2003). A meta–analysis of the criterion–related validity of assessment center dimensions. *Personnel Psychology, 56*, 125–154.

Arthur, W. Jr., & Doverspike, D. (2003). Implications of research and theory for human resource management practices: Selection, recruitment, training, and development. In R. L. Dipboye, & A. Colella (Eds), *Psychological and organizational bases of discrimination at work*. San Francisco: Jossey–Bass [In press].

Arthur, W., Jr., Edwards, B. D., & Barrett, G. V. (2002). *Multiple–choice and constructed response tests of ability: Subgroup performance differences on alternative paper–and–pencil test formats*. [Manuscript submitted for publication].

Arthur, W., Jr., Tubré, T. C., Paul, D. S., & Sanchez–Ku, M. L. (1999). College–sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment, 17*, 354–361.

Arthur, W., Jr., Woehr, D. J., & Graziano, W. G. (2001). Personality testing in employment settings: Problems and issues in the application of typical selection practices. *Personnel Review, 30*, 657–676.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695–716.

Azibo, D. A. (1991). An empirical test of the fundamental postulates of an African personality metatheory. *Western Journal of Black Studies, 15*, 183–195.

Bajtelsmit, J. W. (October, 1975). *Development and validation of an adult measure of secondary cue–using strategies on objective examinations: The test of obscure knowledge (TOOK)*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Ellenville, New York.

Bandura, A. (1997). *Self–efficacy: The exercise of control*. New York: Freeman.

Bangert–Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53*, 571–585.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta–analysis. *Personnel Psychology, 44*, 1–26.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: LEA.

Benson, J., Urman, H., & Hocevar, D. (1986). Effects of test–wiseness training and ethnicity on achievement of third and fifth–grade students. *Measurement and Education in Counseling and Development, 18* (4), 154–162.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta–analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561–589.

Brown, J., Bennet, J., & Hanna, G. (1993). *The Nelson–Denny Reading Test*. New York: Wiley.

Bruch, M. A. (1981). Relationship of test–taking strategies to test anxiety and performance: Toward a task analysis of examination behavior. *Cognitive Therapy and Research, 5*, 41–56.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.),

*Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Cascio, W. F. (1998). *Applied psychology in human resources management.* (5th ed.). Upper Saddle River, NJ: Prentice–Hall, Inc.

Cassady, J. C. (2001). Self–reported GPA and SAT: A methodological note. *Practical Assessment, Research, & Evaluation, 7 (12)*. Retrieved May 21, 2003, from http://ericae.net/pare/getvn.asp?v=7&n=12

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27*, 703–722.

Ceci, S. J., & Williams, W. M. (1997). Schooling, intelligence, and income. *American Psychologist, 52*, 1051–1058.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*, 311–320.

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

Chan, D., Schmitt, N., DeShon, P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationship between race, test performance, face validity perceptions, and test–taking motivation. *Journal of Applied Psychology, 82*, 300–310.

Chan, D., Schmitt, N., Sacco, J. M., & Deshon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*, 471–485.

Civil Rights Act of 1991, Pub. L. No. 102–166, 105 Stat. 1071 (Nov. 21, 1991).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Collins, J. M., & Gleaves, D. H. (1998). Race, job applicants, and the five–factor model of personality: Implications for Black Psychology, industrial/organizational psychology, and the five–factor theory. *Journal of Applied Psychology, 83*, 531–544.

Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement, 34*, 35–46.

Crosby, F. J., Iyer, A., Clayton, S., & Downing, R. A. (2003). Affirmative action: Psychological data and the policy debates. *American Psychologist, 58*, 93–115.

Diamond, J., Ayres, J., Fishman, R., & Green, P. (1976). Are inner city children test–wise? *Journal of Educational Measurement, 46*, 619–625.

Dolly, J. P., & Williams, K. S. (1986). Using test–taking strategies to maximize multiple–choice test scores. *Educational and Psychological Measurement, 46*, 619–625.

Doverspike, D., & Arthur, W. E., Jr. (1995). Race and sex differences in reactions to a simulated selection decision involving race–based affirmative action. *Journal of Black Psychology, 21*, 181–200.

Doverspike, D., Taylor, M. A., & Arthur, W. E., Jr. (1999). *Affirmative action: A psychological perspective*. New York: Nova Science Publishers, Inc.

Eddy, A. S. (1988). *The relationship between the Tacit Knowledge Inventory for Managers and the Armed Services Vocational Aptitude Battery.* Unpublished master's thesis, St. Mary's University, San Antonio, TX.

Educational Testing Service. (2001a). *The college board: The new SAT*. Retrieved July 25th, 2002, from http://www.collegeboard.com/about/newsat/newsat.html

Educational Testing Service. (2001b). *Graduate record examinations: The general test*. Retrieved July 25th, 2002, from http://www.gre.org/stuwrit.html

Ellis, A. P. J. & Ryan, A. M. (1999, April). *Race and cognitive ability test performance: The mediating effects of test–taking strategy use, test preparation, and test–taking self–efficacy*. Paper presented at Society for Industrial and Organizational Psychology conference, Atlanta, GA.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, Department of Justice (1978). Adoption by four agencies of uniform guidelines on employee selection procedures. *Federal Register, 43,* (166), 38290–38315.

Fagley, N. S. (1987). Positional response bias in multiple–choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology, 79*, 95–97.

Fischer, C., Houte, M., Jankowski, M. S., Lucas, S. R., Swidler, A., & Voss, K. (1996). *Inequality by design: Cracking the bell curve myth.* Princeton, NJ: Princeton University Press.

Gibb, B. (1964). *Testwiseness as secondary cue response* (doctoral dissertation, Stanford University). Ann Arbor, MI: University Microfilms (No. 64–7643).

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694–734.

Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*, 691–701.

Goldstein, H. W., Braverman, E. P., & Chung, B. (1992, May). *Method versus content: The effects of different testing methodologies on subgroup differences*. Paper presented at the 7th Annual Conference of the Society for Industrial and Organizational Psychology, Montreal, Quebec, Canada.

Graduate Management Admission Council. (2002). *Analytical writing assessment*. Retrieved July 25th, 2002, from http://www.gmac.com/gmat/the_test/analytical_writing.shtml

Guion, R. M. (1998). Jumping the gun at the starting gate: When fads become trends and trends become traditions. In Hakel, M. D. (Ed.). *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Harmon, M. G., Morse, D. T., & Morse, L W. (1996). Confirmatory factor analysis of the Gibb Experimental Test of Testwiseness. *Educational and Psychological Measurement, 56*, 276–286.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the general aptitude test battery.* Washington, DC: National Academy Press.

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

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47*, 1083–1101.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78–79.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Hoffman, C. C., & Thornton, G. C., III. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50*, 455–470.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions. *American Psychologist, 51*, 469–477.

Hopwood v. State of Texas, 78 F. 3d 932, 948 (5th Cir. 1996).

Horvarth, M., Ryan, A. M., & Stierwalt, S. L. (2000). The influence of explanations for selection test use, outcome favorability, and self–efficacy on test–taker perceptions. *Organizational Behavior and Human Decision Processes, 83*, 310–330.

Hough, L. M. (1992). The "Big Five" personality variables–construct confusion: Description versus prediction. *Human Performance, 5*, 139–155.

Hough, L. M. (1998). Personality at work. In Hakel, M. D. (Ed.). *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future—remembering the past. *Annual Review of Psychology, 51*, 631–664.

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

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology, 83*, 179–189.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86*, 721–735.

Hurtz, G. M., & Donovan, J. J. (2001). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–879.

Jensen, A. R. (1985). The nature of Black–White differences on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences, 8,* 193–263.

Kalechstein, P., Kalechstein, M., & Doctor, R. (1981). The effects of instruction on test–taking skills in second grade African American children. *Measurement and Evaluation in Guidance, 13*, 198–202.

LaFromboise, T., Coleman, H. L. K., & Gerton, J. (1993). Psychological impact of biculturalism: Evidence and theory. *Psychological Bulletin, 114*, 395–412.

Lo, M., & Slakter, M. J. (1973). Risk–taking and test–wiseness of Chinese students. *The Journal of Experimental Education, 42*, 56–59.

Lounsbury, J. W., Bobrow, W., & Jensen, J. B. (1989). Attitudes toward employment testing: Scale development, correlates, and "known–group" validation. *Professional Psychology: Research and Practice, 20*, 340–349.

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology, 47*, 715–738.

Messick, S. M., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191–216.

Miller, P. M., Fagley, N. S., & Lane, D. S., Jr. (1988). Stability of the Gibb (1964) Experimental Test of Testwiseness. *Educational and Psychological Measurement, 48*, 1123–1127.

Miller, P. M., Fuqua, D. R., & Fagley, N. S. (1990). Factor structure of the Gibb Experimental Test of Testwiseness. *Educational and Psychological Measurement, 50*, 203–208.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test wiseness. *Educational and Psychological Measurement, 25*, 707–726.

Millman, J., & Setijadi, A. (1966). A comparison of performance of American and Indonesian students on three types of test items. *Journal of Educational Research, 59*, 273–275.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77–101.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta–analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679–703.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1996, April). *Group differences on overt integrity tests and related personality variables: Implications for adverse impact and test construction.* Paper presented at the 11th annual conference for the Society for Industrial and Organizational Psychology, San Diego, CA.

Outtz, J. L. (1998). Testing medium, validity, and test performance. In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 41–58). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12*, 24–30.

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

Raven, J. C., Raven, J., & Court, J. H. (1994). *A manual for Raven's Progressive Matrices and Vocabulary Scales.* London: H. K. Lewis.

Reynolds, C., & Brown, R. (1984). *Perspectives on bias in mental testing.* New York: Plenum.

Richman–Hirsch, W. L., Olson–Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880–887.

Rogers, W. T., & Yang, P. (1996). Testwiseness: Its nature and application. *European Journal of Psychological Assessment, 12*, 247–259.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*, 537–560.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta–analysis. *Personnel Psychology, 54*, 297–330.

Roth, P. L., Bobko, P., Switzer, F. S., III, Dean, M. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. *Personnel Psychology, 54*, 591–617.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., & Bobko, P. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87*, 369–376.

Rushton, J. P. (2000). *Race, evolution, & behavior: A life history perspective*. (2nd Ed.). Port Huron, MI: Charles Darwin Research Institute.

Rushton, J. P., & Ankey, C. D. (1996). Brain size and cognitive ability: Correlations with age, sex, social class, and race. *Psychonomic Bulletin and Review, 3*, 21–36.

Ryan, A. M. ( 2001). Explaining the Black–White test score gap: The role of test perceptions. *Human Performance, 14*, 45–75.

Ryan, A. M., & Greguras, G. J. (1998). Life is not multiple choice: Reactions to the alternatives. In Hakel, M. D. (Ed.). *Beyond multiple–choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565–606.

Ryan, A. M., Ployhart, R. E., & Friedal, L. A. (1998). Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*, 298–307.

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, D. S. (2000). Applicant self–selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology, 85*, 163–179.

Sacco, J. M., Scheu, C. R., Ryan, A. M., Schmitt, N., Schmidt, D. B., & Rogg, K. L. (April, 2000). *Reading level as a predictor of subgroup differences and validities of situational judgment tests.* Paper presented at the annual conference for the Society for Industrial and Organizational Psychology, New Orleans, LA.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi–predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707–721.

Sackett, P. R., & Roth, P. L. (1996). Multi–stage selection strategies: A Monte–Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49*, 549–572.

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

Sackett, P. R., & Wanek, J. E. (1996). New developments in the use of measures of honesty, integrity, conscientiousness, dependability, trustworthiness, and reliability for personnel selection. *Personnel Psychology, 49*, 787–829.

Sackett, P. R., & Wilk, S. L. (1994). Within–group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*, 929–954.

Sadri, G., & Roberstson, I. T. (1993). Self–efficacy and work–related behavior: A review and meta–analysis. *Applied Psychology: An International Review, 42*, 139–152.

Samson, G. (1985). Effects of training in test–taking skills on achievement test performance: A quantitative synthesis. *Journal of Educational Research, 78*, 261–266.

Sarason, I. G. (1980). Introduction to the study of test anxiety. In I. G. Sarason (Ed.), *Test anxiety: Theory, research, and applications* (pp. 3–14). Hillsdale, NJ: Erlbaum Associates, Inc.

Sarnacki, R. E. (1979). An examination of testwiseness in the cognitive test domain. *Review of Educational Research, 49*, 252–279.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper–and–pencil trades and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30*, 187–197.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–273.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33*, 705–724.

Schmit, M. J., & Ryan, A. M. (1992). Test–taking dispositions: A missing link? *Journal of Applied Psychology, 77*, 629–637.

Schmit, M. J., & Ryan, A. M. (1997). Applicant withdrawal: The role of test–taking attitudes and racial differences. *Personnel Psychology, 50*, 855–876.

Schmitt, N. (1989). Fairness in employment selection. In Smith, M. & Robertson, In. T. (Ed.) *Advances in selection and assessment* (pp. 133–153). New York, NY: John Wiley & Sons.

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

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

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*, 719–730.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46*, 49–76.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp. 290–312). San Francisco: Jossey–Bass.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613–629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African–Americans. *Journal of Personality and Social Psychology, 69*, 797–811.

Steele, C. M., & Aronson, J. (1998). Stereotype threat and the test performance of academically successful African–Americans. In Jencks, C., & Meredith, P. *The Black–White test score gap* (pp. 401–427). Washington DC: Brookings Institution.

Sternberg, R. J., & Hedlund, J. (2002). Practical intelligence, *g*, and work psychology. *Human Performance, 15*, 143–160.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvarth, J. A. (1995). Testing common sense. *American Psychologist, 50*, 912–926.

Taylor–Carter, M. A., Doverspike, D., & Cook, K. (1995). Understanding resistance to sex and race–based affirmative action: A review of research findings. *Human Resources Management Review, 5*, 129–157.

Terris, W., & Jones, J. (1982). Psychological factors related to employees' theft in the convenience store industry. *Psychological Reports, 51*, 1219–1238.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple–choice and constructed response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29–44). Hillsdale, NJ: LEA.

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

Woodley, K. K. (April, 1973). *Test–wiseness program development and evaluation*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wu, T. H., & Slakter, M. J. (1978). Risk–taking and test–wiseness of Chinese students by grade level and residence area. *Journal of Educational Research, 71*, 167–170.

Zuckerman, M. (1990). Some dubious premises in research and theory on racial differences: Scientific, social, and ethical issues. *American Psychologist, 45*, 1297–1303.

# APPENDIX A

## Experimental Test of Testwiseness

Please do not turn this page of the test booklet until directed to do so.

PLEASE DO NOT WRITE ON OR OTHERWISE MARK THIS TEST BOOKLET.

There are a total of 20 items on this test and you have a total of 10 minutes to complete ALL items. In each of the following items, select the alternative that you think is the **BEST** answer and record your choice by filling in the appropriate space on the scantron. Solve as many items as you can. Continue working until you have answered all of the questions or until you are told to stop. Your score is based on the number of *correct* responses. Since there is no penalty for incorrect answers, it is to your advantage to attempt every question.

NOTE: The items are located on the front and back of each page.

1. The Locarno Pact:

   A. is an international agreement for the maintenance of peace through the guarantee of national boundaries of France, Germany, Italy, Belgium and other countries of western Europe.
   B. allowed France to occupy the Ruhr Valley.
   C. provided for the dismemberment of Austria–Hungary.
   D. provided for the protection of Red Cross bases during war times.

2. Henry George is associated with the:

   A. growth of the general disaffection of the working man.
   B. development of the beauties of the great unsettled areas.
   C. settlement of land all across the western part of the country.
   D. the Single Tax program.

3. One trait which the Xerxes of Aeschylus and the Croesus of Herodotus' Solon episode have in common is their:

   A. overconfidence in their own power and wealth.
   B. racial prejudice.
   C. imperialism.
   D. superstitiousness.

4. The career of Marius (155–86 B.C.), the opponent of Sulla, is significant in Roman history because:

   A. he gave many outstanding dinners and entertainments for royalty.
   B. he succeeded in arming the gladiators.
   C. he showed that the civil authority could be thrust aside by the military.
   D. he made it possible for the popular party to conduct party rallies outside the city of Rome.

5. Horace in his 16$^{th}$ Epode typifies:

   A. the despair of the average man when confronted by sweeping social change and its consequent social strains.
   B. the optimism of the average Roman nobleman.
   C. the joy of the common people.
   D. the despair of the political moderate.

6. The first systematic attempt to establish the Alexandrian synthesis between Christian religious belief and Greek civilization was undertaken at:

   A. Rome.
   B. Alexandria.
   C. Athens.
   D. Jerusalem.

7. In Roman times industry failed to attract science to its service because the:

   A. Romans never understood Greek science.
   B. Greeks and Romans despised industry as unworthy of citizens and fee men.
   C. Romans continually destroyed Greek scientific institutions, especially the Museum at Alexandria.
   D. scientific method was completely unknown to the ancients.

8. The industrial revolution in Germany had its greatest momentum:

   A. in the years following the unification of Germany, i.e., from 1871 to 1890.
   B. after the first World War, when Germany, through the loss of territory, was forced to turn to industrialization on a large scale to make a living.
   C. in the 18th century, when several of her absolutist states, adopting mercantilist policies, were pushing industrialization very vigorously.
   D. before the war of 1870–71 in which industrial superiority over France was the main factor responsible for German victory.

9. The liberalizing tendencies of Czar Alexander II came to an end with the:

   A. Civil War in America.
   B. Polish insurrection of 1863.
   C. signing of the Treaty of Calcutta in 1898.
   D. radicalism of the first directorate in France.

10. The expressions "Cavaliers" and "Roundheads" came into use during the struggle between:

   A. Lincoln and his congress.
   B. Charles I and Parliament.
   C. the American colonies and Canada.
   D. the Protestants and the Jews.

11. The Directory:

A. governed France from 1795 to 1799.
B. was overthrown in the 19th century.
C. was the only honest government France has had.
D. was established after the Crimean War.

12. The Locarno Pact:

A. was an agreement between Greece and Turkey.
B. gave the Tyrol to Italy.
C. was a conspiracy to blow up the League of Nations' building at Locarno.
D. guaranteed the boundary arrangements in western Europe.

13. The Progressive Party in 1912:

A. favored complete protective tariffs.
B. favored an appointed Congress.
C. favored the creation of a non–partisan tariff commission.
D. favored restriction of the ballot to certain influential persons.

14. The leading cause of the panic of 1873 was:

A. the deteriorated moral fiber of American Protestants.
B. the assassination of President Lincoln.
C. the overbuilding of railroads.
D. the corruption of the Mormon ruling hierarchy.

15. The Webster–Ashburton Treaty settled a long–standing dispute between Great Britain and the United States over:

A. the Maine boundary.
B. numerous contested claims to property as well as many other sources of ill–will.
C. damages growing out of the War of 1812 and subsequent events.
D. fishing rights on the Great Lakes and in international waters.

16. The Bland–Allison act:

A. made all other forms of money redeemable in gold.
B. standardized all gold dollars in terms of silver and copper.
C. made none of the paper money redeemable in silver.
D. directed the Treasury department to purchase a certain amount of silver bullion each month.

17. Roman imperialism affected landholding by:

A. increasing the number of small farms.
B. eliminating farming in favor of importing all food stuffs.
C. bringing about a more democratic division of land.
D. increasing the number of large estates and reducing the number of small farms thereby increasing the number of landless persons.

18. Charles I and the British Parliament, in their struggles for dominance, developed supporting factions called:

A. Reds and Whites.
B. Cavaliers and Roundheads.
C. Lancastrians and Yorkists.
D. Royalists and Populists.

19. The first great written code of laws was:

A. the Twelve Tables.
B. Corpus Juris Civilis.
C. Hammurabi's Codus Juris.
D. Draco's Juris Categorus.

20. In Roman history, a very famous political controversy developed around the relative power of the civil as opposed to the military components of government. Sulla defended the position that the civil authority was supreme. His opponent, who favored military authority, was:

A. Marius.
B. Cicero.
C. Phidias.
D. Polybius.

# APPENDIX B

## Post–Test Questionnaire

**DIRECTIONS**
We would like some feedback concerning your overall level of motivation to perform well in the present study and your experience with taking standardized achievement tests. Standardized achievement tests are tests for which norms on a reference group, ordinarily drawn from many schools or communities, are provided. Examples of standardized achievement tests are the Stanford Achievement Test and Texas Assessment of Academic Skills (TAAS), and college admissions tests like the Scholastic Aptitude Test (SAT) and American College Testing exam (ACT).

Read each of the questions below and write in or bubble the circle on the accompanying scale that corresponds to your answer.

1. How motivated were you to do your best on each of the tests in this study (i.e., Test Attitudes and Perceptions Survey, Achievement Test, Advanced Progressive Matricies, Reading Comprehension, and History tests)?

   Not at all motivated
   Somewhat motivated
   Quite a bit motivated
   Very motivated
   Extremely motivated

2. How many **MULTIPLE–CHOICE** standardized achievement tests such as the ones described above have you taken within the last four (4) years?

   None
   1
   2
   3
   4
   5
   More than 5

3. Preparation courses are any courses that instruct participants on strategies for taking standardized achievement tests such as the SAT or ACT. Examples of common preparation courses are the Kaplan and Princeton Reviews.

How many test–taking preparation courses have you taken?

| | |
|---|---|
| None | 3 |
| 1 | 4 |
| 2 | More than 4 |

4. How much experience have you had taking **WRITE–IN** or **FILL–IN–THE–BLANK** tests within the last four (4) years?

Not at all experienced
Somewhat experienced
Quite a bit experienced
Very experienced
Extremely experienced

5. What is your current cumulative grade point average (GPA) at your current university? Please round up or down as necessary if your exact GPA falls between an interval (e.g., a 3.47 would fall in the interval of 3.5–4.0).

| | |
|---|---|
| 3.5–4.0 | 1.5–1.9 |
| 3.0–3.4 | 1.0–1.4 |
| 2.5–2.9 | Below 1.0 |
| 2.0–2.4 | I do not yet have a GPA at my current university (e.g., 1st semester freshman) |

6. What was your cumulative grade point average (GPA) in high school? Please round up or down as necessary if your exact GPA falls between an interval (e.g., a 3.47 would fall in the interval of 3.5–4.0).

| | |
|---|---|
| 3.5–4.0 | 1.5–1.9 |
| 3.0–3.4 | 1.0–1.4 |
| 2.5–2.9 | Below 1.0 |
| 2.0–2.4 | |

7. What was your approximate graduating high school percentile rank?

| | |
|---|---|
| $1^{st}$ – $5^{th}$ percentile | $21^{st}$ – $25^{th}$ percentile |
| $6^{th}$ – $10^{th}$ percentile | $26^{th}$ – $30^{th}$ percentile |
| $11^{th}$ – $15^{th}$ percentile | Below the $30^{th}$ percentile |
| $16^{th}$ – $20^{th}$ percentile | |

8. What was your best overall SAT (or ACT) score? _____

9. Please estimate how many items (out of 20) on the Achievement Test that you expect to get correct? _____

**STEREOTYPE THREAT**

**DIRECTIONS**

The following questions ask you to rate your general attitudes and perceptions about standardized tests and your test–taking abilities. Read each of the questions below and check or bubble the circle on the accompanying scale that corresponds to your answer.

| **Strongly Disagree** | **Disagree** | **Neither Disagree or Agree** | **Agree** | **Strongly Agree** |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

| | | |
|---|---|---|
| 1. | I feel confident about my abilities. | |
| 2. | I feel self–conscious when taking tests. | |
| 3. | Tests are often used as a way to discriminate against racial minority group members (i.e., African–Americans). | |
| 4. | I do not feel tests should be used to select people into college or for jobs. | |
| 5. | African–Americans typically do worse than Whites on standardized tests. | |
| 6. | I am a good test–taker. | |
| 7. | I expect to be among the top 20 scorers on the test I completed in this study and therefore expect to win the $30 reward | |

# APPENDIX C

## Test Attitudes and Perceptions Survey

**DIRECTIONS**

The following questions ask you to rate your general attitudes and perceptions about the test you are about to take. Your attitudes should be based on the information in the instruction set read aloud to you and the sample items you just completed. Read each of the questions below and check or bubble the circle on the accompanying scale that corresponds to your answer.

| **Strongly Disagree** | **Disagree** | **Neither Disagree or Agree** | **Agree** | **Strongly Agree** |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

| | | |
|---|---|---|
| *Motivation* | | |
| 1. | Doing well on this test is important to me. | |
| 2. | I want to do well on this test. | |
| 3. | I will try my best on this test. | |
| 4. | I will try to do the very best I can do on this test. | |
| 5. | While taking this test, I will concentrate and try to do well | |
| 6. | I want to be among the top scorers on this test. | |
| 7. | I will push myself to work hard on this test. | |
| 8. | I am extremely motivated to do well on this test. | |
| 9. | I just don't care how I do on this test. | |
| 10. | I won't put much effort into this test. | |
| *Belief in Tests* | | |
| 11. | How well a person scores on this test is a good indication of how well they will do in college classes. | |
| 12. | Tests are a good way of selecting people for college. | |

| Strongly Disagree | Disagree | Neither Disagree or Agree | Agree | Strongly Agree |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

| | | |
|---|---|---|
| 13. | This kind of test should be eliminated. | |
| 14. | I don't believe that tests are valid. | |
| *Anxiety* | | |
| 15. | I probably won't do as well as most of the other people who take this test. | |
| 16. | I am not good at taking tests. | |
| 17. | During a test, I often think about how poorly I am doing | |
| 18. | I usually get very anxious about taking tests. | |
| 19. | I usually do pretty well on tests. | |
| 20. | I expect to be among the people who score really well on this test. | |
| 21. | My test scores don't usually reflect my true abilities. | |
| 22. | I very much dislike taking tests of this type. | |
| 23. | For this test, I find myself thinking of the consequences of failing. | |
| 24. | During a test, I get so nervous I can't do as well as I should have. | |
| *Self–Efficacy* | | |
| 25. | I believe I will perform well on this test. | |
| 26. | I am generally good at performing well on standardized tests. | |
| 27. | Compared with other participants in this study, I expect to do well on this test. | |

| Strongly Disagree | Disagree | Neither Disagree or Agree | Agree | Strongly Agree |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

| | | |
|---|---|---|
| *Face Validity* | | |
| 28. | I do not understand what this test has to do with work required in college courses | |
| 29. | I cannot see any relationship between this test and what is required in college courses. | |
| 30. | It would be obvious to anyone that this test is related to college performance. | |
| 31. | The actual content of this test is clearly related to the work required in college courses. | |
| 32. | There is no real connection between the test that I'm about to take and college courses | |
| *Perceived Predictive Validity* | | |
| 33. | Failing to pass this test clearly indicates that you can't pass many college courses. | |
| 34. | I am confident that this test can predict how well a student will perform in college courses | |
| 35. | My performance on this test is a good indicator of my ability to perform well in college courses. | |
| 36. | Students who perform well on this type of test are more likely to do well in college courses than students who perform poorly on this type of test. | |
| 37. | College admissions counselors could tell a lot about a student's ability to perform in college based on the results of this test. | |
| *Perceived Fairness* | | |
| 38. | The test results will accurately reflect how well I performed on this test. | |
| 39. | I deserve the test results that I will receive on this test. | |

| **Strongly Disagree** | **Disagree** | **Neither Disagree or Agree** | **Agree** | **Strongly Agree** |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 |

| | | |
|---|---|---|
| 40. | This test will fairly reflect my ability to perform well in college courses. | |
| 41. | Overall, I believe that this test is fair. | |
| 42. | I feel good about the way this test will be conducted and administered. | |

# APPENDIX D

## Demographic Form

**DIRECTIONS**
The following questions will provide us with demographic information concerning the participants that complete these measures. None of the information will be used to identify you personally. All data will be analyzed by group averages and not by individual responses. Read each of the questions below and write in your answer or bubble the circle on the accompanying scale that corresponds to your answer.

1. What is your age in years? _____

2. What is your race?

| | |
|---|---|
| American Indian | Hispanic |
| Asian | White |
| African American | Other |

3. What is your sex?

Male
Female

4. What is your college classification?

| | |
|---|---|
| Freshman | Senior |
| Sophomore | Graduate Student |
| Junior | |

5. What is your current cumulative grade point average (GPA) at your current university? Please round up or down as necessary if your exact GPA falls between an interval (e.g., a 3.47 would fall in the interval of 3.5–4.0).

| | |
|---|---|
| 3.5–4.0 | 1.5–1.9 |
| 3.0–3.4 | 1.0–1.4 |
| 2.5–2.9 | Below 1.0 |
| 2.0–2.4 | I do not yet have a GPA at my current university (e.g., 1st semester freshman) |

6. What was your cumulative grade point average (GPA) in high school? Please round up or down as necessary if your exact GPA falls between an interval (e.g., a 3.47 would fall in the interval of 3.5–4.0).

| | |
|---|---|
| 3.5–4.0 | 1.5–1.9 |
| 3.0–3.4 | 1.0–1.4 |
| 2.5–2.9 | Below 1.0 |
| 2.0–2.4 | |

7. What was your approximate graduating high school percentile rank?

| | |
|---|---|
| $1^{st}$ – $5^{th}$ percentile | $21^{st}$ – $25^{th}$ percentile |
| $6^{th}$ – $10^{th}$ percentile | $26^{th}$ – $30^{th}$ percentile |
| $11^{th}$ – $15^{th}$ percentile | Below the $30^{th}$ percentile |
| $16^{th}$ – $20^{th}$ percentile | |

8. What was your best overall SAT (or ACT) score? _____

# VITA

BRYAN DEARA EDWARDS
205 Little Farms Ave.
River Ridge, Louisiana 70123

Doctor of Philosophy, Psychology, August 2003
Texas A&M University
College Station, Texas

Master of Science, Psychology, May 1997
University of South Alabama
Mobile, Alabama

Bachelor of Science, Psychology, May 1995
The University of Alabama
Tuscaloosa, Alabama