Chapter 3

CONSTRUCT-IRRELEVANT VARIANCE IN ACHIEVEMENT TEST SCORES: A SOCIAL COGNITIVE PERSPECTIVE

David E. Ferrier and Benjamin J. Lovett*
Elmira College
Alexander H. Jordan
Dartmouth College

ABSTRACT

Standardized achievement testing is increasingly common in educational and industrial settings. K-12 students take state assessments to comply with federal education laws. Many colleges administer assessments to place incoming students in initial courses and ensure that graduates have benefited from instruction. Professions such as law and medicine give assessments for certification and licensure. Even many employers assess job applicants' levels of literacy and mathematical skills.

Although many of these tests have shown substantial evidence of reliability and validity, there are a host of factors affecting scores on the tests that do not derive from examinees' actual achievement levels. In this chapter, we consider three such factors, reviewing the research on their

* Send any correspondence concerning this manuscript to: Benjamin J. Lovett, Department of Psychology, Elmira College, 1 Park Place, Elmira, NY 14901, e-mail: blovett@elmira.edu

effects and discussing implications for achievement testing in various settings. First, we examine the influence of test-taking motivation on performance. Examinees' motivation to do well on a test varies widely, with predictable effects on resulting scores. We then move on to the impact of test anxiety, including anxiety caused by knowledge of stereotypes concerning one's own group (stereotype threat). Although anxiety is associated with increased motivation to do well, it ironically impedes performance, attenuating estimates of examinees' skills. Finally, we cover the role of prior task exposure and test-taking strategy use on scores. Given the current surfeit of available test preparation materials and services, it is important to understand which types of preparation are actually helpful for examinees.

Throughout the chapter, we note the implications of research on these factors for current achievement testing programs. We also discuss the role of these factors in accounting for group differences in test scores. By the end of the chapter, readers should have a rich understanding of the contributors to achievement test scores (other than achievement itself).

INTRODUCTION

At all levels of education, the standardized assessment of student achievement is increasing in visibility and importance. In K-12 education, students complete teacher-made classroom tests as well as a number of district- and state-wide tests (Wright, 2008). Colleges and universities are also being asked more and more to show that students are meeting "learning outcomes" before graduating (e.g., Dwyer, Millett, and Payne, 2006). Beyond college, achievement tests are used in graduate and professional schools, and in many employment settings as well.

The most important feature of a test is its validity, and in rough terms, a test is said to be valid if it measures what it claims to measure. To state this in more technical terms, when a test is valid, the construct (ability, skill, trait, or domain of knowledge) that it is designed to measure is the source of examinees' scores on the test. One impediment to validity, then, is when factors having nothing to do with the target construct affect examinees' scores. For instance, if the items on a biology examination were printed in very small type, examinees' scores would vary due to variability in visual acuity, and the test would be measuring visual acuity rather than just biology knowledge. This problem, known as "construct-irrelevant variance" (Haladyna and Downing, 2004), is the subject of this chapter.

Specifically, we examine the influence of three factors (other than actual academic skill levels) on achievement test scores. First, we consider how an examinee's interest in and effort on the test can affect his or her scores. Second, we review the effect of test anxiety and related conditions on achievement testing. Finally, we detail how prior exposure to test items or other relevant preparation and training can affect scores. All of these factors have the potential to yield construct-irrelevant variance in achievement test scores, threatening the validity of these tests.

TEST MOTIVATION

When evaluating students' performance scores on achievement tests, it is important to understand when someone might be trying their hardest and when they might not be putting forth the same effort. The difference between these motivational states has a long history of investigation in assessment research; if an individual's score on a cognitive ability test depicts their best effort, this is described as a "maximum performance," whereas if an individual puts forth only an average amount of effort, this performance is labeled "typical" (Sackett, Zedeck, and Fogli, 1988). Although this conceptual distinction is intuitive, in practice, it can be difficult to determine whether an examinee is exhibiting maximal or typical performance. Sackett (2007) argued that if three conditions were met, then one could be confident that scores were indicative of maximum performance. First, the examinee must be aware that they are being evaluated. Second, they must be aware of and accepting of instructions to maximize effort. Third, the performance should occur over a short enough duration that the examinee is able to sustain attention and effort. Of course, these are sufficient rather than necessary conditions; maximum performance may occur over long periods of time, without instructions to maximize effort, and even when examinees do not know that their performance is being evaluated.

To what degree do variations in motivation affect real-world test scores, in practice? One area in which researchers have tried to answer this question is intelligence testing. Since IQ scores have been associated with various life outcomes such as academic and job success (see Jensen, 1998, for a review), it would be both theoretically interesting and practically useful to know if the IQ–life outcome relationship is actually due to differences in motivation levels present when examinees complete the IQ tests—that is, variations in the degree to which examinees show typical versus maximum performance. To

explore the effect of motivation on intelligence test scores, Duckworth et al. (2009) conducted a meta-analysis of previous studies assessing the effects of material incentives (i.e., cash) on intelligence test performance. Across 25 studies of a total of 2,008 participants, material incentives did in fact raise IQ scores with a medium-to-large effect size, Hedges's g = 0.64. Interestingly, material incentives raised the scores of individuals who had baseline IQs (without an incentive) below average more than the scores of individuals whose baseline IQs were above average, g = 0.94 and 0.26, respectively. This led Duckworth et al. (2009) to suggest that since material incentives affected individuals with above-average IQ scores motivation less than their below-average counterparts, those with IQ scores above-average were already closer to maximal performance than their peers.

Of course, if motivation affects IQ scores, the correlation between IQ and life outcomes may itself be due to motivation. This issue was examined in the second part of Duckworth et al.'s (2009) paper. The investigators used data from the Pittsburgh Youth Study, in which 10-year-old boys (approximately 50% of whom were labeled as at-risk due to prior disruptive problems in the school setting) were randomly selected from public schools and followed for three years, during which their teachers gave detailed reports every three months about each boy's performance. When the subjects were 12 years old, they completed intelligence tests. The testing session was videotaped, and later three trained experts coded the children's observable motivation levels. In follow-up interviews conducted when the participants were in young adulthood (average age was 24), the participants were asked to report their years of education, current employment, 12-month history of unemployment, and the number of times they had been arrested.

Duckworth et al. (2009) constructed three multiple regression models for each of the outcomes (performance in school, as well as the long-term life outcomes), while controlling for demographics. One model was fit for IQ scores, the second model for the test motivation values, and the third model utilized both the IQ scores and test motivation values. When looking at how the participant performed academically as an adolescent (i.e., based on the teacher reports of school performance), both IQ and test motivation were each significant predictors when the other was not used, βs = 0.71 and 0.27 respectively. When both were entered in the same regression model, both were still significant predictors, βs = 0.68 and 0.12 respectively. Together, IQ and test motivation were also significant predictors of cumulative years of education, current employment, and number of arrests. Duckworth et al. (2009) also found that test motivation was partially responsible for the

relationships between IQ and school performance, cumulative education, and employment. That is, the relationship between IQ and school performance was attenuated when motivation was statistically controlled. Finally, in agreement with the finding from their meta-analysis showing that material incentives were more effective in increasing IQ scores in participants with below-average IQ, Duckworth et al. found that participants who had below-average IQ scores were also rated lower in test motivation than their above-average counterparts.

If motivation affects consequential real-world test scores as much as Duckworth et al. (2009) suggest, then researchers and applied assessment professionals should consider its influence when interpreting scores. However, it is difficult to know how to apply this to interpreting the scores of individual examinees. One way to do so would involve measuring each examinee's motivation with regard to the test at hand, but as of now, there are few scales developed to do this. However, the industrial psychology literature does report on the development of such a scale, and although it was used in a personnel selection context, we would argue that educational assessment scholars should model this instrument.

Arvey, Strickland, Drauden, and Martin (1990) designed a scale to assess the motivational and attitudinal dispositions of test takers after they had just completed a test in an employment context. A collection of 60 items about test-taking were administered to approximately 500 people and subjected to a factor analysis, which yielded a nine-factor solution, with the nine factors labeled as follows: motivation, lack of concentration, belief in tests, comparative anxiety, test ease, external attributions, general need for achievement, future effects, and preparation. Fifteen items were removed, leaving a 45-item instrument called the Test Attitude Survey or TAS. Initial validation data showed that the TAS yielded expected score differences between examinees taking an easy vs. a difficult test, as well as between job applicants vs. current job holders.

The development of the TAS shows the possibility of assessing examinees' motivation as an aid to interpreting their test scores. Although there are some analogous measures in the educational literature (cf. Wise and DeMars, 2005), more research is needed; equally important, available measures of motivation should be used far more frequently than they currently are.

EXAMINATION-RELATED ANXIETY

The positive association between motivation and performance is an intuitively obvious, easy-to-understand phenomenon. The effect of anxiety on achievement scores is more complex, since examination-related anxiety often stems from the desire to achieve high scores, but it can ironically thwart examinees in reaching that goal. In this section, we review the literature on test anxiety and its effects on performance. In addition, we consider a related phenomenon, stereotype threat, which appears to affect performance by increasing anxiety.

Defining and Measuring Test Anxiety

Although it is relatively common for students to describe themselves as having test anxiety, the meaning of the term is not entirely clear. Putwain (2008) suggests that test anxiety is best understood in relation to general anxiety; the former occurs only in evaluative situations or contexts in which an individual fears performing poorly on tests because a poor score might hinder their academic goals or lead to ostracism from peers. Even this definition does not really clarify the nature of test anxiety, though; at least three possibilities exist. Some literature portrays test anxiety as a stable personality trait, whereas other literature conceptualizes it as an emotional state, referring to the temporary anxiety an individual feels when in an evaluative situation. Trait test anxiety may be one of the determinants of state test anxiety, but the two are distinct, and state test anxiety may occur even when someone's trait test anxiety is low (Zeidner and Matthews, 2005).

A third way that test anxiety is conceptualized is as a clinical disorder, either as a form of *generalized anxiety disorder* (Sapp, Durand, and Farrell, 1995), or as a form of *social phobia* if the test anxiety is extreme enough (McDonald, 2001; Zuriff, 1997).

Consistent with the diverse conceptualizations of test anxiety, many instruments have been designed to measure it. Early measures, such as the Test Anxiety Questionnaire (TAQ) and Test Anxiety Scale (TAS), spawned newer instruments, such as the Test Anxiety Inventory (TAI), Worry Emotionality Questionnaire (WEQ), and Children's Test Anxiety Scale (CTAS). In addition, some researchers stress that test anxiety should be evaluated by qualitative interviews as well as standardized scales (Putwain, 2007).

Measurement and theory development have gone hand in hand, with factor analyses of test anxiety instruments suggesting various facets of test anxiety, and theories of the nature of test anxiety leading to even more recent measures.

The Causes and Effects of Test Anxiety

Early psychodynamic and behavioral theories proposed that test anxiety was caused by such factors as a desire to please one's parents or repeated failures in different areas of life. Unfortunately, there is little research available to support or refute these ideas. Contemporary models emphasize the interaction between personal and situational factors, suggesting that test anxiety is caused by the interaction of personality traits (e.g., one's level of general trait anxiety) and the situation's characteristics (Endler and Parker, 1992; Lazarus, 1999).

One situational factor studied in some detail is known as the *big-fish-little-pond* effect: Students compare their own academic attainments with the attainments of their reference groups (e.g., peers) and use this relative impression as a basis for forming their self-perceptions and coming to conclusions about academic and social status; test anxiety occurs when the students move to a larger reference group and realize that they are no longer the best at what they do (Zeidner and Schleyer, 1999). Several personal factors have also been studied as correlates of test anxiety. In a meta-analysis conducted by Ray Hembree (1988), studies correlating test anxiety with a variety of traits were reviewed. The meta-analysis found a negative correlation between test anxiety and self-esteem (r = -0.42), but a positive correlation between test anxiety and need for achievement (r = .37). In addition, test anxiety was likely to be highest for those individuals who have low ability levels and lowest for people with high ability levels. Demographic factors had small but generally consistent relationships with test anxiety in Hembree's (1988) meta-analysis. First, males had lower test anxiety than females, on average. This result is in accordance with gender stereotypes about anxiety generally (e.g., Swim, 1994). Although Black students showed higher levels of test anxiety than White students in elementary school, these differences diminished across time, becoming virtually nonexistent by the high school grades. Finally, younger siblings generally reported higher levels of test anxiety than siblings who were the oldest in their families.

These correlates and causes suggest that test anxiety is real and predictable, but more relevant for the present chapter is the question of test anxiety's effects on performance. Briefly, these effects are substantial. Hembree's (1988) meta-analysis reported that studies comparing high and low test-anxious examinees found that high test anxiety was related to GPA (r = -.46), performance on standardized tasks of problem-solving and memory (rs = -.45 and -.40, respectively), and the amount of time that examinees took to finish tests (r = .30).

Can the effect of anxiety on performance be diminished? Zeidner (1998, 2007) described various moderating variables influencing the effect of (trait) test anxiety on performance (probably via effects on state test anxiety). If reassurance is provided to examinees, or if the test is highly structured with clear instructions, or if examinees have higher levels of general social support, trait test anxiety has less of an impact on performance. Other ways of minimizing test anxiety's effects include providing examinees with external memory aids containing the information covered on the exam, giving examinees additional easy questions as a way to boost confidence, and providing examinees with information about what the test will cover in advance of actual testing. On the other hand, if the evaluative aspect of the test situation is emphasized, if the test is administered under rigid timed conditions, or if examinees are given negative feedback during the course of the exam, test anxiety will have an enhanced (negative) effect on performance.

Can test anxiety ever help performance? As we noted earlier, there are theoretical reasons to think that it could, since anxiety could lead to studying harder for a test, or it could at least lead to considering test items more carefully and eliminating careless errors. Some theorists refer to "facilitating test anxiety," a construct which encompasses these phenomena (e.g., Mandler and Sarason, 1952). Hembree's (1988) meta-analysis did find a modest positive correlation between facilitating test anxiety and performance on aptitude/achievement tests (r = .29), although more work on this construct remains to be done. Specifically, more attention needs to be paid to the measurement of facilitating test anxiety. Currently, it is measured by self-report questionnaires that ask examinees a variety of items similar to "I do best on exams if I'm nervous." We still do not know whether examinees' self-perceptions regarding facilitating test anxiety are correct—that is, whether state test anxiety actually does lead to improved test performance in some examinees.

A final issue to consider in examining the effects of general test anxiety on performance involves the timing of anxiety measurement. In an intriguing

study, Zeidner (1991) measured students' test anxiety either before or after taking the Scholastic Aptitude Test (SAT). Among those students whose anxiety was measured before the SAT, the anxiety-performance relationship was very small (r = -.11), whereas it was substantially stronger when anxiety was measured after the SAT (r = -.40). Unfortunately, there is no available review of the literature comparing studies that measured anxiety before testing and after testing. However, Zeidner's study should prompt more work on this point.

A Special Case of Text Anxiety: Stereotype Threat

Being aware of negative cultural stereotypes regarding the performance of specific demographic subgroups can have a negative impact on the performance of an individual who is a member of the criticized group. Although this is not a typical case of test anxiety, individuals suffer from anxiety when they worry too much about trying not to fulfill group stereotypes, and ironically, the distracting effects of the anxiety may lead them to perform worse on tests, thus fulfilling the stereotype (Steele et al., 2002). This phenomenon is called stereotype threat and was first described by social psychologists Claude Steele and Joshua Aronson (1995). Literature on stereotype threat has started to accumulate, and a great deal is known about its effects on performance (see Jordan and Lovett, 2007, for a review).

In the original Steele and Aronson (1995) study, two groups of students answered difficult items taken from the Graduate Record Examination (GRE) verbal section. The first group was told that the test they were about to take was indicative of students' verbal ability, while the second group was told that performance on the test had nothing to do with one's ability. Black students who were told that the test was indicative of ability scored significantly worse than White students in the same condition, whereas in the other (non-threatening) condition, White and Black participants achieved comparable scores. That the Black–White performance difference was contingent upon the subjects' belief that their ability levels were being tested was interpreted as evidence for the existence of stereotype threat; activation of examinees' knowledge of racial stereotypes apparently impaired test-taking ability.

Stereotype threat effects have been found for a variety of groups. In addition to the aforementioned effects among Black test-takers (Blascovich, Spencer, Quinn, and Steele, 2001; Steele and Aronson, 1995), researchers have documented intellectual performance decrements due to activation of

negative cultural stereotypes attached to Latinos (Gonzales, Blanton, and Williams, 2002) and low-SES individuals (Croizet and Claire, 1998). Women, too, have been found to suffer in their performance on math tests when stereotypes concerning gender and math ability are made salient (Swim, 1994; Davies, Spencer, Quinn, and Gerhardstein, 2002). Moreover, as a means of protecting themselves against these stereotype threat effects, some individuals whose abilities run counter to cultural stereotypes may distance themselves from their demographic groups, possibly losing part of their heritage or individuality in the process (Fryer, 2006; Steele and Aronson, 1995; Zirkel 2004). For example, women who are exceptionally gifted at math may downplay their femininity to avoid thinking about the stereotype regarding women and mathematical ability (Pronin, Steele, and Ross, 2004).

While the mechanics behind stereotype threat are still not entirely understood, anxiety appears to be the most likely mechanism. Initial empirical work cast doubt on this explanation, since experimental participants' self-reports of anxiety did not statistically mediate stereotype threat effects (Aronson et al., 1999). However, in more recent studies such as Blascovich et al. (2001), participants' blood pressure was significantly higher in situations containing stereotype threat while participants' blood pressure was actually lower than normal in non-threatening conditions. Other studies have shown that trained experts' ratings of anxiety (based on participants' fidgeting) as well as participants' decreased heart-rate variability mediate stereotype effects (Croizet et al., 2004; Bosson, Haymovitz, and Pinel, 2004). The discrepancy between self-reports and other measures of anxiety is intriguing, and suggests that participants may not always be aware of their stereotype-related anxieties, perhaps interpreting their heightened physiological arousal as a different emotion (e.g., excitement). In any case, several studies suggest that stereotype threat is another example of how anxiety can lower test performance, leading to underestimates of students' skills.

TEST SOPHISTICATION

One way to reduce the effect of anxiety on performance is to give examinees information about the test that they are going to take; information generally comforts examinees, giving them a chance to mentally prepare for the task that they are about to begin. Giving examinees information about exams has effects far beyond anxiety reduction, though; test sophistication, a term that we will use broadly to refer to any provision of information or skills

in an effort to increase future test performance, often *does* increase performance. In a classic article on this topic, Anastasi (1981) distinguished three sophistication strategies leading to improved test scores: training in general cognitive skills, training in specific test-related skills, and exposure to the actual test. As Anastasi notes, training in general cognitive skills (i.e., training to increase intelligence) has been advocated by a number of psychologists since the time of Alfred Binet, whose work on "mental orthopedics" (Binet, 1911) was designed to improve the cognitive skills of intellectually deficient school children. Admittedly, the efficacy of such exercises is hotly disputed, and there is literature to support both sides of the debate (Brown, 1978; Whimbley, 1975, 1980; Bloom, 1976). Supporters claim that increases in intelligence can even occur at later ages, although most research concludes that early training is most effective (Anastasi, 1981).

Exposure to a test is a very different kind of sophistication, but it has similar effects on test performance, generally increasing it. Sometimes these effects are known as practice effects, and test scores typically tend to be higher for those who are taking a test a second time (Angoff, 1971; Droege, 1966; Powers and Camara, 1999). Although this is a potential threat to the validity of tests (if scores indicated how many times examinees were exposed to the testing situation, rather than their actual academic skills), test sophistication effects can be equated across examinees to some degree through the application of brief orientation and practice sessions (Wahlstrom and Boersman, 1968). Of course, these sessions will not equate for variability in prior exposure to actual test items, but for tests such as the SAT, where students take a different version each time that they take the test, orientation and practice sessions would at least equate for exposure to the general testing experience.

Test Preparation Effects

Most test sophistication, though, does not consist of exposure to the actual test items, nor does it consist of training in general cognitive skills. Instead, most test sophistication involves training in rather specific test-related skills, something often described as test preparation or coaching. This is seen most vividly in commercial programs that brag about their ability to increase scores on college admissions tests and related measures. While many test preparation providers claim that their programs will improve an individual's score significantly, this is often an exaggeration. Claims of 100-point gains on the

SAT, for example, might be closer to actual improvements of only approximately 30 points, a gain equal to about three correct answers on the SAT (Briggs, 2009).

This seemingly small gain in test performance is still enough to encourage many students to prepare for admissions tests in ways that they might not ordinarily for other tests. Many of these students, typically in their junior or senior year of high school, feel the pressure to take the SAT or ACT (research has shown that the percentage of high school seniors taking admission tests increased from 51% in 1992 to 63% in 2004; see Briggs, 2002; Briggs and Domingue, 2009), and if these students' performance was not satisfactory to themselves or the institutions to which they were seeking admission, then they would have two options: look elsewhere for schools or improve their test scores. Briggs (2009), on behalf of the National Association for College Admission Counseling (NACAC), distributed hundreds of surveys to NACAC-member colleges, asking about their admissions procedures. One of the findings was that a respectable number of colleges and universities (over *one-third* of the institutions surveyed) said that a slight 20-point gain on the SAT could "significantly improve a student's likelihood of admission" (Briggs, 2009, p. 17). Obviously, this would depend on where a student's score starts out; a rise from 200 to 220 on the SAT-Verbal is unlikely to impress colleges. Still, the colleges' responses to Briggs's survey suggest that students may be justified in seeking even small score gains.

In reviewing the literature on coaching effects, Briggs (2009) found evidence for several points. First, there was a positive effect of coaching on SAT performance, although the effect was of small magnitude, as mentioned earlier. Second, coaching often led to larger improvements on math sections of the SAT than on critical reading (e.g., Powers and Rock, 1999); indeed, it was difficult to find evidence that any coaching programs substantially increase SAT verbal performance. Third, although the percentage of students taking admissions test has increased, the purchasing of test preparing books, commercial coaching services, tutoring, and other forms of coaching generally remained the same between 1988 and 2002 (Briggs, 2002; Briggs and Domingue, 2009).

Overall, it should be made clear that although some forms of coaching or test preparation are likely to increase scores in college admission tests, there is no set rule as to which type of test preparation, if any, is appropriate for any one person. Moreover, an individual's motivation to do well on high-stakes tests is likely to moderate any effects of test preparation services. In addition, there is still little research currently available on the efficacy of online test

preparation services, despite their recent popularity. Finally, ideal research on the assessment of test preparation effects would involve the random assignment of subjects into different test preparation conditions, and as of 2009, no research has incorporated such methodology.

These coaching studies all involved college admissions tests, but it is important to note that test preparation is also used for the high-stakes accountability tests so prevalent today in K–12 schools. Recent federal legislation, such as the No Child Left Behind (NCLB) Act, has resulted in more coordinated statewide use of standardized achievement tests, and decisions about school funding and other important issues are tied to school performance. Test preparation is one strategy used by states and school districts to ensure that sufficient numbers of students are performing at adequate levels (Lai and Waltman, 2008).

Test-Wiseness: A General Type of Sophistication

A final sophistication-related variable—one not mentioned by Anastasi (1981)—that may explain variability in achievement test scores is test-wiseness. A classic definition of test-wiseness is "a subject's capacity to utilize the characteristics and formats of the test and/or test-taking situation to receive a high score" (Millman, Bishop, and Ebel, 1965, p. 707). A test-wise examinee may show a number of traits outlined in a classic taxonomy developed by Millman and colleagues. Six traits were presented in this taxonomy, four of which apply to any test. First, test-wise examinees use time-management strategies (e.g. going over answers on a test with any remaining time to reconsider/ensure confidence in responses). Second, test-wise examinees use a strategy for minimizing mistakes (e.g. making sure that the directions are clearly read and understood before answering a question). Third, test-wise examinees employ a guessing strategy (e.g. if there is no penalty for an incorrect answer, then never leave a question unanswered). Finally, test-wise examinees use deductive reasoning (e.g. getting rid of answer choices that are known to be incorrect and choosing from the remaining options). The other two traits depended on the test's constructor and/or the purpose of the test. First, test-wise test-takers pay attention to the test constructor's advice as to what will be on the exam (e.g. pay attention to the relevance of certain details). Lastly, test-wise examinees use cues (e.g. the correct answer is typically longer or shorter than the incorrect options).

Test-wiseness can be a threat to the validity of scores unless it is held constant across examinees (Rogers and Yang, 1996). To be held constant would require one of two things: either test-wiseness effects are present for everyone or they are absent for everyone. To ensure the former possibility, efforts could be made to raise the test-wiseness of individuals with low test-wiseness. Training programs could be incorporated into the curriculum of middle schools and continued up to the high school level (Samson, 1985; Sarnacki, 1979). Alternatively, to reduce test-wiseness effects across all examinees, the responsibility lies mainly on the test constructor. Contradictory distracters, grammatical incongruence between distracters and item stems, and leaving correct answers substantially longer or shorter in length than distracters are the sorts of cues that test-wise examinees rely on to bolster their scores. These cues could be at least partially eliminated through training of the test constructors (Rogers and Yang, 1996).

CONCLUSION

Motivation, anxiety, and sophistication are all associated with variability in performance on standardized cognitive tests, across a large number of studies, measures, and populations. The degree to which examinees have an incentive to perform well, feel nervous about their performance, and are knowledgeable about the tests they are about to take can all have a significant impact on the resulting scores on achievement tests. The evidence suggests that motivation, anxiety, and sophistication are not just potential sources of construct-irrelevant variance, but actual sources.

Although some scholars and assessment practitioners may find this conclusion depressing, we prefer to view it as a challenge. One way to eliminate sources of construct-irrelevant variance is to keep them from varying across examinees. For instance, we discussed above how test-wiseness effects could be diminished if all examinees were trained to be test-wise. This is a more general solution to the problems discussed in this chapter. The research on motivation should challenge test users (and researchers) to maximize motivation for all examinees, by providing effort-based incentives, designing briefer, more engaging test stimuli, and building rapport before asking examinees to put forth their best efforts. Similarly, the research on test anxiety should lead to widespread efforts to screen students for evaluative anxiety so that intervention can occur early; we have effective treatments (Zeidner, 1998) and no excuse for not using them more widely.

We conclude on a brief theoretical note. Recent approaches to measurement validity (e.g., Kane, 2006; Messick, 1995) argue that it is not tests themselves that are valid, but uses of tests and inferences made on the basis of test scores. The research reviewed in the present chapter should be taken as suggesting that achievement tests should be used cautiously when motivation, anxiety, or sophistication are likely to vary across examinees. Moreover, achievement test scores may at times lead, in the context of other information, to valid inferences about examinees' levels of motivation, anxiety, and sophistication, rather than their academic skills. For instance, when a student who has always done very well in school does poorly on an achievement test, the score may be evidence of the student's anxiety, or lack of preparation, or low motivation. All of this complicates the interpretation of achievement test scores, but as we have argued, it is to be hoped that this realization will lead ultimately to more motivated, less anxious, and increasingly test-wise examinees, hardly a consequence to be derided.

REFERENCES

Aronson, J., Lustina, M. J., Good, C., Koeogh, K., Steele, C. M., and Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35*, 29-46.

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

Blascovich, J., Spencer, S. J., Quinn, D. M., and Steele, C. M. (2001). Stereotype threat and the cardiovascular reactivity of African Americans. *Psychological Science, 12*, 225-229.

Bosson, J. K., Haymovitz, E. L., and Pinel, E. C. (2004). When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety. *Journal of Experimental Social Psychology, 40*, 247-255.

Briggs, D. C. (2002). SAT coaching, bias and causal inference. *Dissertation Abstracts International*. DAI-A 64/12, p. 4433. (UMI No. 3115515).

Briggs, D. C. (2009). *Preparation for College Admission Exams*. National Association for College Admission Counseling.

Briggs, D. C. and Domingue, B. W. (2009). The effect of admission test preparation: new evidence from ELS:02. Unpublished Working Paper. www.colorado.edu/education/faculty/derekbriggs/publications.html.

Croizet, J. C., and Claire, T. (1998). Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin, 24*, 588-594.

Croizet, J. C., Despres, G., Gauzins, M. E., Huguet, P., Leyens, J. P., and Meot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin, 30*, 721-731.

Davies, P. G., Spencer, S. J., Quinn, D. M., and Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin, 28*, 1615-1628.

Duckworth, A. L., Quinn, P. D., Lynam, D., Loeber, R., Stouthamer-Loeber, M., Moffit, T. E., and Caspi, A. (2009). *What intelligence tests test: Individual differences in test motivation and IQ*. Manuscript submitted for publication.

Dwyer, C. A., Millett, C. M., and Payne, D. G. (2006). *A culture of evidence: Postsecondary assessment and learning outcomes*. Princeton, NJ: Educational Testing Service.

Endler, N. S. and Parker, J. (1992). Interactionism revisited: Reflections on the continuing crisis in the personality area. *European Journal of Psychology, 6*, 177-198.

Fryer, R. G. (2006, Winter). "Acting white": The social price paid by the best and the brightest minority students. *Education Next, 6*(1), 53-59.

Gonzales, P. M., Blanton, H., and Williams, K. J. (2002). The effects of stereotype threat and double-minority status on the test performance of Latino women. *Personality and Social Psychology Bulletin, 28*, 659-670.

Haladyna, T. M., and Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*, 47-77.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger Publishers/Greenwood Publishing Group.

Jordan, A. H., and Lovett, B. J. (2007). Stereotype threat and test performance: A primer for school psychologists. *Journal of School Psychology, 45*, 45-59.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: ACE/Praeger.

Lai, E. R. and Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice, 27*, 28-45.

Lazarus, R. S. (1999). Stress and emotion: A new synthesis. New York: Springer.

Mandler, G., and Sarason, S. (1952). A study of anxiety and learning. *Journal of Abnormal and Social Psychology, 47*, 166-173.

McDonald, A. (2001). The prevalence and effects of test anxiety in school children. *Educational Psychology, 21*, 89-101.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

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

Powers, D. E., and Camara, W. J. (1999). *Coaching and the SAT I: College Board Research Note 06.* New York: College Board.

Powers, D. E. and Rock, D. A. (1999). SAT Coaching: What really happens and how we are led to expect more. *The Journal of College Admission, 129*, 7-17.

Pronin, E., Steele, C. M., and Ross, L. (2004). Identity bifurcation in response to stereotype threat: Women and mathematics. *Journal of Experimental Social Psychology, 40*, 152-168.

Putwain, D. W. (2007). Researching stress and anxiety in schoolchildren: Some methodological considerations. *British Educational Research Journal, 33*, 205-217.

Putwain, D. W. (2008). Deconstructing test anxiety. *Emotional and Behavioural Difficulties, 13*, 141-155.

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

Sackett, P. R., Zedeck, S., and Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482-486.

Sackett, P. R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance, 20*, 179-185.

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

Sapp, M., Durand, H., and Farrell, W. (1995). The effects of mathematics, reading and writing tests in producing worry and emotionality test anxiety with economically and educationally disadvantaged students. *College Student Journal, 29,* 122-125.

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

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. *Journal of Educational Measurement, 19,* 211-219.

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

Steele, C. M., Spencer, S. J., and Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 34* (pp.379-440). New York: Academic Press.

Swim, J. K. (1994). Perceived versus meta-analytic effect sizes: An assessment of the accuracy of gender stereotypes. *Journal of Personality and Social Psychology, 66,* 21-36.

Wise, S. L., and DeMars, C.E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1-17.

Wright, R. J. (2008). *Educational assessment: Tests and measurements in an age of accountability*. Thousand Oaks, CA: Sage.

Zeidner, M. (1991). Test anxiety and aptitude test performance in an actual college admission testing situation: Temporal considerations. *Personality and Individual Differences, 12,* 101-109.

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum.

Zeidner, M. (2007). Test anxiety in educational contexts: Concepts, findings, and future directions. *Emotion in Education,* 165-184.

Zeidner, M. and Matthews, G. (2005). Evaluation anxiety. In A. J. Elliot and C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 141-163). London: Guilford Press.

Zeidner, M. and Schleyer, E. (1999). The big-fish-little-pond effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology, 24,* 305-329.

Zirkel, S. (2004). What will you think of me? Racial integration, peer relationships and achievement among white students and students of color. *Journal of Social Issues, 60*, 57-74.

Zuriff, G. E. (1997). Accommodations for test anxiety under ADA? *Journal of American Psychiatry and Law, 25*, 197-206.