From *Measuring Noncognitive Variables for Student Success and Retention: Improving Admissions and Student Services for Diverse Groups Including Women and Students of Color*
By William Sedlacek

The following is an excerpt from the unedited manuscript.

Chapter Two

"Tradition is the prison where change is detained..."

Israelmore Ayivor

Traditional Admissions Measures

Should a student come to higher education fully developed, or do we wish to select on dimensions on which a student will improve through experience at an institution? There has been a recent focus on "college readiness" suggesting information separate from the wide range of attributes a student will need once enrolled (Conley, 2005). While readiness for college includes taking the appropriate courses, getting good grades, and scoring well on admissions tests, there is evidence that many other attributes determine whether most students will succeed in higher education (Sedlacek, 2011).

Courses

Sedlacek (2011) has suggested that while students continue to need courses in math, English, foreign languages, etc., there has been a tendency among educators and college admissions staff to feel that more is better. The reasoning goes that if we would just require more

courses in certain areas (eg math), students would be better prepared. However, the law of diminishing marginal utility from economics becomes relevant at some point (Diamond & Rothschild 1989). The logic behind the law is that beyond a certain level, there is little or no increase in the value of more units in a given area. Thus, at some point, number of courses in a subject or field may no longer be relevant as a predictor. We may have reached an asymptote, and the variable has become a constant.

For example, Sawyer (2008) studied 245,175 students from 9,507 high schools who took the EXPLORE (8th grade), PLAN (10th grade), and ACT (12th grade) tests. He concluded that taking additional standard college preparatory courses in high school, taking advanced/honors courses, and earning higher grades would, by themselves, only modestly increase the percentage of students who leave high school adequately prepared to take credit-bearing courses in the first year of college. Sawyer also felt that taking additional courses and earning higher grades mostly benefit students who by grade eight are already well on a their way to getting ready for higher education. He concluded that developmental variables also should be considered.

In summary, up to a point, more math and other courses are useful in preparing students for higher education. Beyond that point, other variables, such as those presented in this book, become more important for student success. In fact, one could argue that an over dependence on some courses may detract from time spent on other important aspects of a student's development.

Grades

Recent literature has shown that grades are becoming increasingly less useful as indicators of student achievement or as predictors of future student success (Sedlacek, 2011). This is largely due to the statistical artifact that students at all levels of education are being assigned higher grades. Are current students just smarter and/or more accomplished than their predecessors? This seems unlikely, but even if true, it does not help us prepare students for higher education, since grades no longer appear as useful in differentiating levels of student academic achievement as they were once.

Grades have become more of a constant because of "grade inflation." For example, Woodruff and Ziomek, (2004) found that the mean grade point average (GPA) of high school students taking the ACT assessment had increased a total of .20 to .26 points on a four-point system from 1991 to 2003, depending on the subject area. Rojstaczer & Healy (2012) showed that the mean GPA in higher education nationally had risen from 2.94 in 1991–1992 to 3.11 in 2006–2007, on a four-point system. Rojstaczer & Healy (2010) concluded that there is a nationwide rise in grades over time of a roughly 0.1 change in GPA per decade. However, they found that relative to other schools, public-commuter and engineering schools tend to grade "harshly", making GPA an even more complex variable of questionable utility in selecting applicants from any group in higher education.

Marquardt (2009) noted that some school districts in Virginia were offering students an increase in their course grades or overall GPA as an incentive to take the Commonwealth's Standards of Learning examination. Marquardt found that the mean GPA of first year students in Virginia colleges and universities rose from 3.27 to 3.56 on a four-point system, between 1995 and 2007, compared to an increase in GPA in a national sample during that same period of 3.28 to 3.49. Additionally nationally, many K–12 schools were not assigning grades to students and were using extramural and portfolio assessments instead (Washor, Arnold & Mojkowski, 2008).

Rojstaczer & Healy (2010) found that private high schools were grading 0.1 to 0.2 higher than public schools on a 4.0 scale for a given talent level of student. Since the evidence indicates that private schools may educate students no better than public schools (Perscarella and Ternzini, 1991), private schools seemed to be giving an advantage to their students by awarding higher grades. They graded easier and there was a tendency for graduate schools, professional schools, and some employers to assume increased competencies for those who have attended selective private schools (Bernstein, 2003; Burrelli, Rapoport & Lehming, 2008). As one Dartmouth faculty member put it "We began systematically to inflate grades, so that our graduates would have more A's to wave around" (Perrin, 1998).

Lahr et al. (2014) found that performance-based funding was increasingly popular among both state and federal policy makers in a survey of college administrators in Indiana, Ohio and Tennessee. These administrators wanted public institutions to graduate more students, more efficiently. However, a common way that colleges dealt with those funding formulas was by using grade inflation or admitting fewer "at-risk" students.

There has been an increase in so called “pathway” programs for international students matriculating in U.S. colleges and universities (Winkle, 2014). Typically, such programs have an extra year or period of taking courses in a special program at a U.S. institution. The program may be affiliated with a for-profit organization and be quite expensive for participants. Winkle studied such programs at 12 institutions and concluded that large numbers of pathway students recruited by the corporate partner were not prepared for college level credit work. Often they were given full academic credit for inflated grades they received in those preparatory courses. The pressure to recruit international students and their related fees has escalated in recent years. The increased income from out of state applicants is particularly attractive to state funded institutions.

I can recall a White female undergraduate student I had in a course on racism. She was uncomfortable with the content and did not participate in class discussions which she was told were part of her grade. She earned a “B” in the class and went to the Dean to complain that she always got “As”, and this would ruin her chances to go to law school. I had a long conversation with her on the fairness of her grade and that one grade would be unlikely to decide her future.

Interviews

Interviews are a common method employed by programs in higher education, particularly in professional schools. Muchinsky (1987) noted that interviews can take many forms from highly structured to open-ended questions that may vary by interviewer. Interviews may be the most difficult method of assessment on which to achieve reliability and validity. Because there are so many variables that need to be addressed and controlled in the interview setting, extensive training of interviewers is necessary. Shaw & Milewski (2004) discussed such things as regular calibration sessions, and using multiple interviewers to achieve reliability and consistency in measurement.

In my experience in training interviewers to assess noncognitive variables, there are several key concepts I use. First, interviewers must be given a rubric to use in scoring. They should feel they have to provide an assessment on certain dimensions before they finish. Second, interviewers should be made aware that their ratings would be evaluated; ideally in open discussion retraining sessions. This makes it easier to get all interviewers using the same

methods. Third, interviewers should be told that they are not after their own general assessment of the candidate, but rather part of a team that is scoring specific variables. This reduces the chances that their own styles and ideas will be a variable. These points should be discussed in training and retraining. Fourth, most interviewers will start following the desired structure, but will "drift" from that structure over time, as they do more interviews, get tired etc. Encouraging interviewers to be fresh, take breaks, work at the best time of day for them if possible etc. is advised. Fifth, interviewers should be made aware of their biases for or against certain candidates. We all have them, regardless of our characteristics or experience. Appendices A1 and A2 provide example scoring systems that can be used in assessing noncognitive variables from interviews.

Whether biases are based on race, gender, religion, sexual orientation, age, country of origin etc, some training on this is very important. Various methods of providing such training are available (Westbrook & Sedlacek, 1988, Sedlacek, 2004a; Sandlin & Sedlacek, 2013; Wilson, Sedlacek & Lowery, 2014). Prieto et al, (1978, 1986) and Prieto & Sedlacek (1990) developed the Simulated Minority Admissions Exercise to train admissions committees in medical schools to interview applicants from a variety of backgrounds, races, and cultures.

Longerbeam, Sedlacek, Balón, & Alimo, (2005) studied the prejudices of professionals working in diversity and multicultural offices at three universities. These professionals were working in all areas of diversity. They found what they called the "Multicultural Myth", in that 70% of participants felt they had no prejudices. The authors also concluded that working in one area of diversity did not free one from prejudice in another. Again, that prejudice could be positive (Halo Effect) or negative. In my training experience I have found people with positive or negative predispositions on such variables as physical size, physical attractiveness, accent, field of study, dress, participation in athletics, cheerleading, and interest in hunting, among others. None of these attributes were relevant for the particular purpose of the interview, but variance due to those dimensions came in, often without the awareness of the interviewer.

Letters of Recommendation

Recommendations suffer from a number of measurement problems (Sedlacek & Prieto, 1990, Sedlacek 2004a). They tend to yield unreliable and invalid results for all groups, often

suffering from a positive Halo Effect. Recommendations tend to give overly positive and undifferentiated comments across candidates; everyone looks great. Dirschl & Adams (2000) found low reliability in multiple ratings of letters of recommendation for a residency at the University of North Carolina School of Medicine. Additionally, they found a low correlation with performance in the residency of the selected applicants.

Aamodt & Williams (2005) studied more than 11,000 letters of recommendation from 51 different studies of students and employees and found that they did not add incremental validity to the combination of GRE scores and undergraduate GPA. They concluded that one reason was that few references were negative and that two studies indicated fewer than 7% of students or job applicants received average or below average reference ratings. This was especially true when students didn't waive their right to see their references. Aamodt & Williams noted that letters of recommendation might differ in the traits or characteristics used to describe an applicant, but will seldom differ in how the evaluator judges the quality of the applicant.

The results of recommendations tend to be more useful if the recommender is known to the evaluator or the evaluator requests the recommendation from a particular person. In order to increase the variability in letters of recommendation and make them more valid and reliable as indicators of student success, writers of letters should be asked to cover specific topics in the letters they write, such as examples of behavior or answers to specific questions. If a recommendation is carefully evaluated as one piece of application evidence, then reliability, in the form of corroborative information, might be established. Without being specific, the letters may be of little use to evaluators (Sternberg 2010). The evaluations of noncognitive skills that ETS, and those studying law school admissions, have focused entirely on assessments through letters of recommendation. Sternberg felt that the efforts of ETS in revising their tests do not go far enough. Colleges and universities could get more useful letters by being clear in asking those submitting letters to focus on the issues that the ETS system evaluates, he said. Strenberg felt that a better approach would be to have applicants themselves submit evidence of their noncognitive skills.

The University of Maryland School of Medicine demonstrated that recommendation letters were an important part of the evaluation of the plaintiff's application to a medical school in defending their use of noncognitive variables in selecting students (*Farme*r v. *Ramsay et al.,* 1998). Because the letters were written by people known to the medical school, they could be

coordinated with other application materials. If there were inconsistencies across different kinds of information presented, candidates could be asked to explain any discrepancies.

In the Gates Millennium Scholars program, letters of recommendation were provided to evaluators, along with grades, activities and personal statements by candidates. This is a scholarship program for candidates of color. Evaluators are trained to consider all the material in assessing applicants. With training and multiple sources of information, evaluators can differentiate among candidates, achieving reliabilities above .90 and validity in predicting success in higher education (Sedlacek & Sheu, 2008). Appendices A1 and A2 provide scoring systems that can be employed with letters of recommendation.

Portfolio Assessment

Portfolios are yet another way to do assessment in higher education (LaMahieu, Gitomer, & Eresch, 1995). In this method, examples of a person's creative work are presented for evaluation. Portfolios have been commonly used in the arts, architecture and design, to demonstrate the work of applicants for admission. Chen and Mazow (2002) have discussed the value of electronic learning portfolios for students to present their accomplishments. Smith & Tillema (2003) identified three important issues in portfolio assessment: The clarity and explicitness of the portfolio collection; the feasibility of the collection process itself; and trust in the outcome of what is being collected.

The school of design at one university has required an additional admissions procedure beyond the general one employed for all undergraduates. Traditionally, it has required a portfolio containing design-related materials produced by the applicant. Administrators and faculty at the school wished to broaden the content of the portfolio to contain information on noncognitive variables, such as how the applicants had overcome obstacles, how they saw themselves, and what their goals were. The school officials felt this would give them better information with which to judge their applicants, particularly those of color. In another example, one potential problem in portfolio assessment; that middle-class students may benefit most, faculty evaluators were trained in identifying examples of high and low scores on noncognitive variables (Koretz, 1993).

The University of California, Irvine, has employed a Personal Achievement Profile along with SAT or ACT scores, grades, and specific courses completed, as part of its admission profile (Wilbur & Bonous-Hammarth, 1998). The University included, among other things, the noncognitive variables of leadership, community service, and creative achievement. After applicants were screened on their academic credentials, about 60 percent of the entering class was determined. The additional 40 percent of the class was selected on the basis of the Personal Achievement Profile. Using a double-blind procedure, admissions staff, who were trained in reviewing the profiles, made the judgments. No interviews or letters of recommendation were employed, and the entering class varied across a number of dimensions.

YES Prep is a program intended to increase higher education attendance and graduation of students from low income commuities.YES Prep has high school students develop a College Assessment Portfolio Project (CAPP) based on the eight noncognitive variables described in Exhibit 1 as one of the Pillars in their student success programs (YES Prep, 2014). Their progress on improving on the noncognitive variables is documented over the four years of high school and monitored and evaluated by counselors and teachers. More information on the YES Prep program is covered in Chapter Eight.

The Big Picture Learning (BPL) curriculum is mostly experiential without grades in typical courses. In order to present their potential for admission to colleges and universities, BPL high school students employ portfolios. Students present information about how they are improving on noncognitive variables in their portfolios as well as other accomplishments and creative work (BPL, 2014). Utilizing different approaches and creating new forms that fit the particular needs of schools or programs is encouraged. This increases the probability that noncognitive variables can be used to benefit students in a variety of contexts. A detailed discussion of BPL's uses of noncognitive variables will be presented in Chapter Eight.

Sternberg (2010) noted that the optional portion of the enrollment application at Tufts University, where he was a Dean, also included other possible formats that could emphasize applicants' strengths outside of traditional application materials. Students could, for example, submit a YouTube video about themselves, use a sheet of paper to create something, blueprint a future home, create a new product, draw a comic strip, design a costume or a theatrical set, compose a score or do something entirely different.

Such procedures are useful techniques to generate the material one would need to assess noncognitive variables. The next critical step would be to have a method that could be used to score those materials in a consistent (reliable) way, along some dimensions that have been shown to have some validity in correlating with performance criteria for students, such as grades, retention, or other attributes of success in higher education.

Essays

Essays have been shown to be very unreliable indicators of applicant potential, unless raters are trained to score them in specific ways and reliability across raters is established (Sedlacek & Prieto, 1990; Sedlacek, 2004a). With appropriate training, it is possible to have raters score essay material on noncognitive variables. For example, in the Gates Millennium Scholars program, readers were able to score essays on the applications with high interjudge agreement (Pearson r = .81) on the noncognitive variables shown in Exhibit 1. The scores obtained from evaluating the essays were added to an overall score that showed a statistically significant positive relationship with grades and retention in higher education (Sedlacek & Sheu, 2008).

Appendix A1 can be employed to help identify how people may provide information on noncognitive variables in an essay. Appendix A2 contains some sample cases, with names and identification information changed, that could be used in training for evaluation of essays.

Application Review

Application review can be seen as similar to essays or portfolios in that an application containing many facets (among them essays and presentations of information on a person) is evaluated. It is important to demonstrate consistency in the ratings of the full application. Shaw & Milewski (2004) called this composite reliability. Applications can be reviewed even if there was no *a priori* intention to evaluate noncognitive variables. For example, in the first year of the Gates Millennium Scholars program, the applications were designed and completed by applicants before a determination was made as to how evaluation of the applications would be done. Despite this, the applications contained information that could be evaluated for

noncognitive variables. Reviewers were trained to score the applications across all the materials presented which included grades, personal statements, letters of recommendation, activities, courses taken, and information on the secondary school attended (see Appendices A1 and A2).

Inter-judge Pearson correlations of between .81 and .85 were achieved using this process in several situations involving admissions and financial aid.

Application review is a good way to start the process of using noncognitive variables in that one can review application materials of current matriculants to do a retrospective study of which noncognitive variables appeared to relate to successes or failures among students. This technique was employed as one of the evaluation methods in the *Farmer* v. *Ramsay et al.* (1998) medical school case which is discussed below.

In summary, noncognitive variables present a method of improving assessments for all students and are particularly useful for nontraditional students. Noncognitive variables can be assessed in a number of ways: questionnaires, interviews, essays, portfolios, and reviewing materials not specifically designed to elicit information on noncognitive variables. In the next chapter, the use of noncognitive variables in admissions and scholarship selection is discussed. Chapter Eight includes many examples of how noncognitive variables have been used by schools and programs in higher education .

Tests

Admission tests were created initially to help select as well as advise students (Sedlacek 2004b). They were intended to be useful to educators making decisions about students. While they were always considered useful in evaluating candidates, tests were also considered to be more equitable than using prior grades because of the variation in quality among preparatory schools. The College Board has long felt that the SAT was limited in what it measured and should not be relied upon as the only measure to judge applicants (Angoff, 1971).

In 1993, the verbal and mathematical reasoning sections of the SAT were lengthened and the multiple-choice Test of Standard Written English was eliminated. The name was changed from Scholastic Aptitude Test to Scholastic Assessment Tests, while retaining the SAT

initials. It is now officially labeled the SAT. In 2003, the College Board announced that an essay would be added and that the analogies item type would be removed as of 2005.

The ACT, first administered in 1959, was presented as an alternative to the SAT, emphasizing a wider range of abilities than simply verbal and math. The paradox might be that SAT and ACT scores have always been highly correlated, despite their apparent intentions over the years. The ACT was intended to be more achievement-based while the SAT was seen as more of an aptitude test (Willingham, Lewis, Morgan, & Ramist, 1990, Sedlacek, 1998a, 2004b). The websites of both programs provide tables to covert scores from one test to the other, thus assuming equivalence (ACT, 2014; College Board, 2014a). The University of Maryland switched from requiring the ACT to requiring the SAT in the late 1960s. Studies there showed the correlation between the two tests to be .88 for University of Maryland applicants. In other words, the two tests were measuring something operationally very similar.

In 2013 the ACT test was administered to 1.8 million people, compared to 1.7 million who took the SAT. Perhaps motivated by this fact, in March, 2014 the College Board announced that major changes to the SAT would be coming in 2016 (College Board, 2014b). The SAT redesign is centered on eight key changes. The revised SAT will focus on relevant words, the meanings of which depend on how they are used. When people take the evidence-based reading and writing section of the new SAT, they will be asked to demonstrate their ability to interpret, synthesize, and use evidence found in a wide range of sources.

An attempt will be made to align the revised SAT with the practical work of college and career. In the essay section on the new SAT, students will read a passage and explain how the author builds an argument. The math section will stress problem solving and data analysis, along with algebra, and advanced math. In the revised SAT, test takers will be presented an excerpt from one of the Founding Documents or a text from the ongoing Great Global Conversation about freedom, justice, and human dignity. The CEO of ACT has been critical of the new SAT- ACT conversion tables.He felt the new SAT scores are inflated and do not seem to be converted fairly (Roorda, 2016).

Whether these changes result in the measurement of different attributes that are more fair and relevant to the potential of a diverse applicant pool remains to be seen. Despite various changes and versions over many years, the SAT in essence still measures what it did in 1926;

verbal and math ability. It is basically still a general intelligence test that does not assess a wide range of abilities (Sedlacek, 2003a, 2004b).

Most studies on the SAT or ACT focus on student performance in the first year of higher education as the criterion. Why don't standardized tests relate to measures of student success beyond the first-year? Aside from not being designed to do so, Sternberg (1996) pointed out that such tests measure only one aspect of intelligence; analytic ability. He defines analytic ability as "one's capacity to interpret information in a well-defined and unchanging context." Sternberg felt standardized tests generally do not measure creative ability or practical ability; the two other components of intelligence he identifies. Persons with creative ability are able to interpret information in changing contexts. They can easily shift from one perspective to another. They are likely to be the best researchers or contributors to their fields. Persons with practical intelligence know how to interpret and use the system or environment to their advantage.

If we examine a typical curriculum, many would agree that creative and practical intelligence come into play more in the later years of most programs, since upper-level courses tend to require students to write more, discuss more, and hopefully think more. Analytic skills, as defined by Sternberg, appear less useful by themselves beyond the first-year. Sternberg (2010) felt that it has been clear for years that traditional measures account for only some of the difference in academic performance of students and that noncognitive variables also are important, so these changes are long overdue. Noncognitive variables appear to be in Sternberg's (1985, 1986, 1996) creative or practical ability areas, while tests tend to reflect analytic ability.

Aptitude Versus Achievement Tests: The Curious Case of the Indestructable Straw Person

The distinction between aptitude and achievement has been a focus of psychometricians for many decades (Jenks & Crouse, 1982). While psychologists can discuss the differences between the two conceptually, they are difficult to sort out empirically. Kelley demonstrated that widely used intelligence tests and achievement batteries overlapped by about 90% (Kelley, 1927). Anastasi (1984) discussed the dilemma of constructing a "straw person" that prompts a question that cannot be definitively answered. She felt that all tests assess current

status, whether their purpose is terminal assessment or prediction. She added that traditional achievement tests often can serve as effective predictors of future learning, since all tests are measures of a sample of behavior.

Both aptitude and achievement tests can be best characterized as tests of developed ability. It may be most useful to categorize measures by their actual function, rather than their labeled or intended function. Anastasi concluded, with regard to cognitive behavior, that test scores tell us what an individual is able to do at a given time. They do not tell us why individuals perform as they do. To answer that question, we need to know something about each person's experiential background (Anastasi, 1984). Here is where noncognitive measures become relevant.

The GRE

The Graduate Record Examination (GRE) also has been criticized for not being related to student success in graduate school. Sternberg & Williams (1997) found correlations in the low teens to .20 between grades in graduate school and the GRE. The strongest relationship was found for the analytical portion of the exam, which no longer exists. They noted the methodological issues of studying only those who enrolled in graduate school and the questionable success criterion of grades for graduate students. Kaplan & Saccuzzo (2009) found that GRE scores correlated highest with first year grades for graduate students, supporting Stenberg & Williams' findings of correlations of .20 or below. Kaplan & Saccuzzo concluded that "the GRE predict[s] neither clinical skill nor even the ability to solve real-world problems" (p. 303).

Graduate and professional school students have a smaller GPA range than undergraduates, making the criterion measure less useful psychometrically. Restriction of range tends to depress relationship statistics, such as correletion coefficients. Graduate and professional students also matriculate in much smaller, more relatively independent programs than do undergraduates, making sampling for research purposes much more difficult. Also, we expect different qualities from our graduate students than from undergraduates, thus the importance of noncognitive variables again becomes relevant (Sedlacek, 1972, 2004,a; Sedlacek & Prieto, 1990).

Lemann (2000) felt that we had come to a point where the "Big Test" had become the primary object of attention in many schools. It had become the standard by which we judge ourselves and others. Many assume that if an individual has high ACT, SAT, or GRE scores, or if a school has high mean scores on such tests, the students must be learning something, and the school must be good. This is ascribing too much power to our tests. Posselt (2016) found that many leading departments, despite saying otherwise, are reluctant to admit anyone who does not have exteremely high GRE scores.

ETS (Educational Testing Service) began offering the PPI (Personal Potential Index) in 2009 as optional for GRE applicants. It is an online system of evaluation designed to provide additional information on graduate school applicants in six areas: Knowledge and Creativity; Communication Skills; Teamwork; Resilience; Planning and Organization; and Ethics and Integrity. ETS studied the validity of this system for evaluating graduate school applicants and its fairness to students of color and other groups. While ETS felt results were encouraging, too few applicants and graduate programs used the PPI. As a result, when ETS planned large-scale validity studies, it didn't have enough results to be statistically valid, and decided to phase out the program (Payne, 2016).

ETS will be offering graduate departments assistance in showing why they can admit applicants from a wider distribution of GRE scores than many currently do (Payne, 2016). If departments provide data on those who are admitted with slightly lower GRE scores than is the norm for their program, Payne believed research would show that a broader range of GRE scores can lead to comparable levels of academic success in doctoral programs (Payne 2016).

While it was good to see ETS continuing to pursue alternative variables and approaches as aids to educators in admissions, one of the difficulties in the ETS approach is that there is an attempt to rule out any variance due to diversity on race, gender, culture, religion and other dimensions. In that variance is the key to understanding the potential of many applicants. It is logical that a testing program would want to have a single scoring system that could be applied to everyone in the same way.

There is evidence that noncognitive variables can be viewed as constructs that could be interpreted in the context of the applicant's race, culture, gender, sexual orientation and other nontraditional dimensions. By doing this for all applicants, including the most traditional, an admissions program can add some unique and useful information that correlates with the success

of students beyond their scores on standardized tests or alternative measuring instruments that do not allow for diversity (Sedlacek,1994b).

Miller & Stassun (2014) in their article "A Test That Fails" argued that the GRE should be deemphasized, and measures of other attributes such as drive, diligence and the willingness to take scientific risks should be added. They felt this would make graduate admissions more predictive of the ability to do well, and would also increase diversity in STEM (science, technology, engineering & math) fields.

Miller & Stassun stated that in the physical sciences, only 26% of women, compared to 73% of men, score above 700 on the GRE Quantitative measure. For "minorities", it was 5.2%, compared with 82% for Whites and Asians. They felt that the "misuse" of GRE scores to select applicants may be responsible for the under-representation of women and minorities in graduate school. Women received 20% of U.S. physical sciences PhDs, and under-represented minorities, who account for 33% of the U.S. university-age population, earned 6%. They noted that these percentages are similar to the percentage of students who score above 700 on the GRE quantitative measure. They concluded that: "In simple terms, the GRE is a better indicator of sex and skin colour than of ability and ultimate success."

A few innovative STEM PhD programs, such as those at the University of South Florida and Fisk–Vanderbilt, have achieved completion rates above 80%, which is well above the national average, and are increasing participation by women and minorities (Powell, 2013). Their admissions process includes an interview that assesses college and research experiences, key relationships, leadership experience, community service, life goals, and perseverence (Miller & Stassun, 2014).

The American Astronomical Society (2016) has called for limiting the use of the GRE in admissions for graduate admissions in the astronomical sciences.

Why Use Tests At All?

Standardized tests remain controversial in general, particularly their fairness for people of color (Sedlacek, 1976, 1977b; FairTest, 2007; Helms, 2009; Toldson, 2014, Guinier, 2015). Much of the debate centers on statistical artifacts, measurement problems, and poor research methodology, including biased samples and inappropriate statistical analyses and interpretations

(Sackett, Borneman, & Connelly, 2009). While this discussion and related controversy is useful and interesting to academics, we may have lost track of why tests were developed to begin with, and their limitations. Test results should be useful to educators, student service workers, and administrators by providing the basis to help students learn better and to analyze their needs. As currently designed, tests do not accomplish these objectives. The newest SAT, or another test, may change that conclusion, but the results are years away. Many teachers tend to teach to get the highest test scores for their students, student service workers may ignore the tests, and too many administrators are satisfied if the mean test scores rise in their schools. We need something from our tests that currently we are not getting. We need measures that are fair to all and provide a good assessment of the developmental and learning needs of students, while being useful in selecting outstanding applicants. Our current tests don't do that.

Test-Optional Programs

Many schools make ACT or SAT tests optional for admissions. FairTest (2014) listed more than 800 schools that did not require tests for admission to bachelor's degree programs. Hiss & Franks (2014) studied Bates College graduates for more than 20 years under an admission test-optional program. Additionally, they researched optional standardized testing policies in the admissions offices at 33 public and private colleges and universities, using cumulative GPA and graduation rates. The researchers also examined which students were more likely to make use of an optional testing policy, and what benefits could derive from such a policy. Four types of institutions were included in the study: 20 private colleges and universities, six public universities, five "minority-serving" institutions and two institutions focused on the arts; for a total of nearly 123,000 students and alumni. The difference in graduation rates between submitters (those who provided test scores) and non-submitters was 0.10 percent. The mean undergraduate GPA of submitters was 3.11 while for non-submitters it was 3.06.

Academic ratings assigned to applicants by Bates admissions staff were equally accurate whether or not test scores were submitted. Under the test-optional policy, application rates increased for students of color and women as well as those for those from low-income and blue-collar backgrounds. The policy also helped students with learning disabilities and international

applicants gain admission. Non-submitters were more likely to major in fields that emphasized creativity and originality.

Belasco, Rosinger & Hearn (2014) studied the relationship between test-optional policy implementation and subsequent growth in the proportion of low-income and minority students enrolling at liberal arts colleges. Additionally, they analyzed whether test-optional policies increased numbers of applications. Results showed that test-optional policies tended to increase the perceived selectivity, rather than the diversity, of the participating institutions. Also, after controlling for a number of variables, they concluded that there was no evidence that test-optional policies increased student applications for admission. Schools that implemented test-optional policies reported an increase in mean test scores, which is consistent with past studies (Ehrenberg, 2002; Yablon, 2001). However, adoption of test-optional policies did not increase the proportion of low-income and African-American, Latino or Native-American students who enrolled.

Making tests optional for admission does not necessarily guarantee a more diverse student body. It is a start, but experienced admissions staff can make judgments that are less efficient substitues for the same variance that tests represent. Staff training on how to evaluate a more diverse pool of potential applicants may be required. DePaul University trains its admission staff to do an assessment of noncognitive variables for 5% of its entering class in a test-optional program (Kalsbeek, 2016). Advice to any institution would be to analyze why a student of color, or students from other groups, would want to attend your institution. Emphasizing variables that relate to campus climate can make a difference to such prospective students. Issues of how racism is handled, courses offered in areas relating to diversity, and student development programs relevant to different student groups may be important.

While the test optional movement is interesting and deemphasizes the use of tests for admissions, we are still left with the question "What will replace tests in our admissions models?" Just because we make tests optional does not mean that we have replaced them with something different. I have done an exercise with a number of admissions committees by presenting applicant materials without test scores, and asking them to estimate what their test scores would be. An experienced admissions person can make a quite accurate guess, because they are used to assessing the types of candidates presented at their institution. Based on variables related to test scores, such as grades, socioeconomic status, secondary school attended,

activities, personal statements, recommendations and other aspects of the application, admissions experts can make those judgments. The point of the exercise is that if we do not use tests, we need to make sure we are not simply assessing a surrogate for those scores. What we can end up with is a lower quality measure of the same variance represented by a test.

Keeping Up With Change

Change does not roll in on the wheels of inevitability, but comes through continuous struggle.

Martin Luther King Jr.

The world is much different than it was when the ACT, SAT, GRE and other tests were developed in the last century. International students, women, people of color, lesbian, gay, bisexual, transgender or questioning (LGBTQ) students, and people with disabilities, among others, are participating in higher education in more extensive and varied ways (Knapp, Kelly, Whitmore, Wu, & Gallego, 2002; Meyer, 2003; Wawrzynski & Sedlacek, 2003, Sedlacek & Sheu, 2004a,b; Longerbeam et al 2005; Kalsbeek, Sandlin, & Sedlacek, 2013). The percentage of U.S. college students who are Hispanic, Asian/Pacific Islander, Black, and American Indian/Alaska Native has been increasing. From 1976 to 2011, the percentage of Hispanic students rose from 4 percent to 14 percent; the percentage of Asian/Pacific Islander students rose from 2 percent to 6 percent; the percentage of Black students rose from 10 percent to 15 percent; and the percentage of Native American/Alaskan Native students rose from 0.7 to 0.9 percent. During the same period, the percentage of White students fell from 84 percent to 61 percent (National Center for Education Statistics, 2013). Commonly employed tests have not kept up with the implications of these changes (Sedlacek, 2011).

It is not good enough to feel constrained by the limitations of our current methods of assessing abilities or potential. Instead of asking, "How can we make our admissions tests better?" we need to ask, "What kinds of measures will meet our needs now and in the future?" In a presentation I made to the American Astronomical Society (Sedlacek, 2014a), I used the analogy of the multiverse (Greene, 2011). As the name suggests, the assumption of the multiverse is that

there are other universes in the Cosmos that are outside our own. In statistics we define a universe as the largest group or domain to which we wish to generalize our results. The idea of the multiverse here is that we need to look to another universe of content to measure variables that are unique to that universe. These other variables may not overlap or share variance with the traditional measures, but may relate to criteria of interest, such as grades and retention. The purpose of Chapter Three is to present the underlying logic and research supporting a method that yields such measures. We do not need to eliminate tests from our admissions systems; we need to add some new measures that expand the range of dimensions we consider. Somewhere in that range is a variable or variables that capture the potential of all comers.