St. John's Law Review

Volume 80
Number 1 *Volume 80, Winter 2006, Number 1*

Article 7

February 2012

How and Why Standardized Tests Systematically Underestimate African-Americans' True Verbal Ability and What to Do About It: Towards the Promotion of Two New Theories with Practical Applications

Dr. Roy Freedle

THE RONALD H. BROWN CENTER FOR CIVIL RIGHTS AND
ECONOMIC DEVELOPMENT
SYMPOSIUM

HOW AND WHY STANDARDIZED TESTS SYSTEMATICALLY UNDERESTIMATE AFRICAN-AMERICANS' TRUE VERBAL ABILITY AND WHAT TO DO ABOUT IT: TOWARDS THE PROMOTION OF TWO NEW THEORIES WITH PRACTICAL APPLICATIONS

DR. ROY FREEDLE†

INTRODUCTION

In this Article, I want to raise a number of issues, both theoretical and practical, concerning the need for a total reassessment of especially the verbal intelligence of minority individuals. The issues to be raised amount to a critical reappraisal of standardized multiple-choice tests of verbal intelligence, such as the Law School Admissions Test ("LSAT"). I want to probe very deeply into why such standardized tests systematically underestimate verbal intelligence.

This leads me first to review the prospects for a new standardized test of verbal intelligence associated with the studies of Joseph Fagan and Cynthia Holland.[1] These studies show us that the races are equal; this result leads us to question the *construct validity* of many current standardized tests of verbal aptitude. Then, I briefly review my own studies of

† Formerly of Educational Testing Service, Princeton, NJ (Retired). The author can be contacted at freedle2@aol.com.

[1] *See* Joseph F. Fagan & Cynthia R. Holland, *Equal Opportunity and Racial Differences in IQ*, 30 INTELLIGENCE 361 (2002).

standardized tests that suggest a systematic underestimation of the ability of minorities. My studies question not only the *construct validity*, but also the *reliability* of the scores used to assess individual test performance, especially for minority students and even White students from lower socio-economic strata. In order to correct some additional problems associated with standardized testing, I present in some detail a new theoretical model to explain the concept of "*guessing*" as it occurs on standardized tests. After assessing the empirical adequacy of this new guessing theory, I then apply it to help clarify why it is absurd to describe the test behavior of especially low scoring students as being due to "lucky guessing" especially when it comes to describing minorities' choices in response to hard test items. Properly assessing guessing behaviors provides us with another means to criticize the *validity* and *reliability* of scores reported for many current standardized tests. Next, I touch on student mentoring issues, the temporal aspects of testings,[2] the structure of law school course work,[3] and how these issues can affect the *predictive validity* of tests such as the LSAT. Finally, I review a central point of contention for this conference: namely, the *validity* of using primarily incoming students' LSAT scores to rank the quality of law school faculty. I will take up each of these validity and reliability issues in turn.

I. FAGAN & HOLLAND: A TEST DEMONSTRATING RACIAL EQUALITY IN VERBAL INTELLIGENCE

I first wish to discuss a new standardized test of verbal intelligence that successfully embraces a culture-free assessment of ethnic ability. The test is due to Fagan and Holland.[4] The Fagan-Holland test is important because it provides us with a standardized test that demonstrates racial equality in verbal intelligence.

Let me repeat the above assertion. The important result in

---

[2] *See* William D. Henderson, *The LSAT, Law School Exams, and Meritocracy: The Surprising and Undertheorized Role of Test-Taking Speed*, 82 TEX. L. REV. 975, 1024–34 (2004).

[3] *See* DONALD E. POWERS & SPENCER S. SWINTON, EDUC. TESTING SERV., EFFECTS OF SELF-STUDY OF TEST FAMILIARIZATION MATERIALS FOR THE ANALYTICAL SECTION OF THE GRE APTITUDE TEST 2 (1982), *available at* http://ftp.ets.org/pub/gre/gre-79-9r.pdf.

[4] *See* Fagan & Holland, *supra* note 1, at 384–85.

the Fagan-Holland test of verbal intelligence is that *no* evidence for ethnic differences in intelligence occurs using their procedures.[5] Some of my earlier work suggested that if one used what I called a revised SAT scoring method, this would lead to a decrease of about one-third in the mean separation of African-Americans and Whites.[6] By contrast, the Fagan-Holland test leads to a total erasure of the mean difference between these two groups.[7] To be sure, there are still strong individual differences to be found—that is, some students always score very high on this test, while others score consistently low—but when mean racial groups are compared, there are NO significant differences. This I consider to be a major breakthrough in assessing verbal intelligence in sharp contrast to the very negative things that, for example, Arthur Jensen has said in the past regarding racial differences in intelligence.[8]

So, here is what I would like to suggest to law school admissions officers. I would like you to consider using the Fagan-Holland test of verbal aptitude either as a replacement for the LSAT or at least in addition to the LSAT. That is, if one really wants to honestly assess an incoming student's verbal aptitude in a manner that is totally free of racial bias, here is the opportunity to do it. I say, use the Fagan-Holland test.

So you must be asking: What is the Fagan-Holland test? And why does it produce the results that it does? For your convenience, the details of their procedures underlying this new test are contained in Table 1.

---

[5] *See id.* at 364–66.

[6] See Roy O. Freedle, *Correcting the SAT's Ethnic and Social-Class Bias: A Method for Reestimating SAT Scores*, 73 HARV. EDUC. REV. 1, 21–22 (2003) [hereinafter Freedle, *Correcting the SAT's*]; Roy O. Freedle, *The Truth and the Truthful Sages That Spin It: A Review of Dorans*, 74 HARV. EDUC. REV. 73, 74–75 (2004) [hereinafter Freedle, *A Review of Dorans*].

[7] *See* Fagan & Holland, *supra* note 1, at 380.

[8] *See* Arthur R. Jensen, *How Much Can We Boost IQ and Scholastic Achievement?*, 39 HARV. EDUC. REV. 1, 4–5 (1969) (arguing that compensatory education efforts should be reexamined because they failed to create positive gains in children's IQ).

Table 1

THE FAGAN-HOLLAND PROCEDURE DEMONSTRATING THE EQUALITY OF THE RACES

The basic outline of the Fagan-Holland studies showing the equality of the races is as follows.

(a) They selected 40 words from the dictionary that were extremely rare–so that neither White nor African-Americans students had ever heard of these words. An example Fagan & Holland gave is the word VENTER. (Have you ever heard of this word? I certainly have not.)

(b) Next, they exposed students of both races to sentences that used each unfamiliar word in an appropriate context. For example, the word "venter" was used in the following sentence: "Tubby had a big, fat venter."

(c) They then asked each student "a simple question about the unknown word [used] in the sentence to see if the person being trained understood the meaning of the sentence. They had to indicate, for example, whether a venter was a body part or a mental state." In other words, people had to INFER the meaning of the rare word from its use in a sentence. Narrowing down the choice to either a "body part" or a "mental state" assisted in carrying out the correct inferential process.

(d) An irrelevant fifteen minute task was then presented to prevent the students from rehearsing these new words.

(e) Finally, all students were tested on how well they knew the meaning of these forty new words within the context of a multiple-choice vocabulary test. For example, they were asked:

(f) Which of the five choices is the meaning of "venter": a) height, b) candle, c) badge, d) belly, e) opening?

Individuals who were good learners and good at drawing inferences would select "belly" as the correct answer.

Let me briefly describe what they did and why they did it. Fagan and Holland basically built a new IQ test—which in their

case was essentially a carefully constructed vocabulary test.[9] Under the assumption that racial differences in intelligence do *not* exist, Fagan and Holland reasoned that *if all individuals are given an equal opportunity to learn all the crucial concepts that are needed in order to select the correct answer* on any standardized test, there then should be *no* significant difference in how the races respond to the test.[10] That is, the mean number of correct items should be statistically equivalent when compared across races or other ethnic groups.

Their results confirmed their hypothesis. When one rigorously controls the amount of experience in learning the relevant concepts underlying a test so that all racial groups get equal exposure to these concepts, there will be no evidence of significant ethnic differences. Fagan and Holland also concluded that conventional IQ tests—which use concepts for which students have *not* had equal opportunities over the years (either in home or at school) to learn the precise meanings and associations of key terms in the test content—can be expected to yield racial differences.[11] And indeed, to demonstrate this last point they also administered to these same individuals, who had just yielded no mean racial differences in IQ, a conventional IQ test and found the old racial differences in mean IQ reappearing.[12] Thus, conventional IQ tests are racially biased in terms of mean correct responses because they make the false assumption that all examinees have had equal opportunity to learn the concepts and materials used in the test, an assumption that is patently false.[13] Incidentally, my own studies of ethnic bias in the Scholastic Aptitude Test ("SAT") also point to *vocabulary* as at least one of the culprits in yielding false conclusions regarding racial and ethnic differences.[14] A similar point has been made by Clifford Hill and Eric Larsen in their provocative studies of how African-American third-graders systematically misinterpret test questions due to differences in use between the races regarding semantic and syntactic cues

---

[9] *See* Fagan & Holland, *supra* note 1, at 365–66.

[10] *See id.* at 364.

[11] *See id.* at 363.

[12] *See id.* at 374–76.

[13] *See id.* at 385. Interestingly, individual students who score high (or low) in the Fagan-Holland IQ test also tend to score high (or low) in the biased IQ test.

[14] *See* Freedle, *Correcting the SAT's*, *supra* note 6, at 28–29.

contained in the formulation of many test questions.[15] For example, one of their vocabulary examples that illustrates semantic confusion across the races involved the use of the word "home."[16] Hill and Larsen point out that "home" has a broader meaning among African-Americans due to the fact that the African-American community involves an "extended family" structure.[17] Therefore, "home" to an African-American third-grader might refer equally to grandma's house, mother's house, or auntie's house. Most White Americans no longer have an extended family structure and so "home" does not have this ambiguous set of referents. Since the test question under consideration assumed a non-extended family structure, African-American children frequently got that item incorrect. In the individual interviews conducted to determine how each item was processed, it was clear that the African-American children knew what the question and its associated reading passage meant. But, nonetheless, the test makers were unwittingly punishing them for belonging to a community with a different social structure. The White children were not similarly punished.[18] Many other ethnic differences on test items covering a wide variety of semantic and syntactic principles were reported by Hill and Larsen.[19] Despite such clear-cut explanations for mean test differences, scholars like Jensen conclude that there is a strong genetic component underlying the observed ethnic differences from observing only the mean ethnic differences in correctly responding to such biased tests. The Fagan and Holland work definitely contradicts such a shallow conclusion.[20]

Therefore, the Fagan-Holland test calls into question the validity and reliability of standardized tests of verbal intelligence.[21] That is, if the Fagan-Holland verbal intelligence test results are correct, and if one takes their findings as a solid demonstration of racial equality in verbal intelligence, then any

---

[15] *See* CLIFFORD HILL & ERIC LARSEN, CHILDREN AND READING TESTS (2000).

[16] *Id.*

[17] *Id.*

[18] *See* C. Steele & J. Aronson, *Stereotype Threat and the Test Performance of Academically Successful African-Americans*, *in* BLACK-WHITE TEST SCORE GAP, 401 (Christopher Jencks & Meredith Phillips eds., 1998); *see also* Jay Rosner, *On White Preferences*, NATION, Apr. 14, 2003, at 24, 24.

[19] *See* HILL & LARSEN, *supra* note 15.

[20] *See* Fagan & Holland, *supra* note 1, at 362–63, 380–82.

[21] *See id.* at 384–85.

other standardized test—such as the LSAT or the SAT or the Graduate Record Examination ("GRE")—that leads to mean ethnic differences MUST therefore be judged to be culturally biased and hence is an invalid measure of minority ability. We will use this important inference in building a bridge to some of the remaining points that will be made below.

Before leaving the topic of the Fagan-Holland test, I would like to draw a clear link to the main reason for this conference: the concern that the use of the LSAT (which by the above inference must be ethnically biased) is strongly influencing the *U.S. News & World Report*'s ranking of law school quality, and leads to law school admissions officers systematically rejecting more and more minority students who are known as a group to score lower on this standardized test. I would like to point out both here and in the conclusion of this Article that if the Fagan-Holland test of verbal aptitude had been used in place of the LSAT, there would be no further need for conferences of this type. That is, since the races are equal when the Fagan-Holland test is used, it would not matter, in terms of racial admission rates, whether the incoming students' aptitude scores were or were not included in the *U.S. News & World Report*'s rankings of law school quality. But the uncomfortable reality is that the LSAT *is* the test that is used to influence the rankings, and the LSAT is, by the above inference, racially biased. So until another test replaces the LSAT, we are stuck with the endless dilemma of how to solve the racial admissions problem with the admissions procedures now in place. Real change is hard for schools but is harder still for minorities. And by 2050, the problem, due to anticipated demographic shifts, will be even more discomfiting.

II. OTHER EARLIER STUDIES OF ETHNIC BIAS IN STANDARDIZED TESTS: A BRIEF REVIEW OF MY STUDIES

Beginning in 1978, my colleagues and I found in various computer analyses that there is a persistent pattern of ethnic bias that occurs on many standardized tests, including eleven SAT forms administered in the 1980s and thirteen of the earlier GRE forms from the same time period (in particular, the paper-and-pencil version of this test).[22] I used a technique called the

[22] *See* ROY FREEDLE & IRENE KOSTIN, EDUC. TESTING SERV., RELATIONSHIP BETWEEN ITEM CHARACTERISTICS AND AN INDEX OF DIFFERENTIAL ITEM

Differential Item Functioning ("DIF") statistic to establish this result, a method invented by Dorans and Kulick.[23] That is, when you match the scores of African-Americans and Whites—say, you pick only people who scored 500 on their verbal SATs—you'll find something surprising. The African-Americans will score *better* than expected on the hardest items and worse than expected on the easiest items;[24] this bias pattern persists across almost the entire ability spectrum from the lowest scores of 200 to very high SAT scores. I used this persistent bias pattern to show how to re-estimate what the SAT scores should have been for at least the African-American students. Each minority student ended up with two SAT scores: the standard SAT score along with the Revised-SAT score. Because one can construct two scores for each minority student with both of them typically different in magnitude, one can question the reliability of the SAT test. Furthermore, because the reliability is questioned, the validity of the ability construct that the test is intended to measure is also questionable. Nevertheless, as already mentioned earlier in this Article, the results for the new Revised-SAT scores indicated that the mean separation between White and Black test performance would be reduced by one-third, which, by conventional standards, is a very large amount. Therefore, it seems reasonable that the adoption of my particular method for rescoring the SATs for at least the minority students (and disadvantaged White students

---

FUNCTIONING (DIF) FOR THE FOUR GRE VERBAL ITEM TYPES 39 (1988), *available at* http://www.ets.org/research/researcher/RR-88-29.html; Roy Freedle & Irene Kostin, *Item Difficulty of Four Verbal Item Types and an Index of Differential Item Functioning for Black and White Examinees*, 27 J. EDUC. MEASUREMENT 329, 332 (1990); Roy Freedle & Irene Kostin, *Predicting Black and White Differential Item Functioning in Verbal Analogy Performance*, 24 INTELLIGENCE 417, 417–18, 425, 442 (1997) [hereinafter Freedle & Kostin, *Predicting Black and White*]; *see also* ROY O. FREEDLE, IRENE W. KOSTIN & LARAINE M. SCHWARTZ, A COMPARISON OF STRATEGIES USED BY BLACK AND WHITE STUDENTS IN SOLVING SAT VERBAL ANALOGIES USING A THINKING ALOUD METHOD AND A MATCH PERCENTAGE-CORRECT DESIGN (1987).

[23] *See* NEIL J. DORANS & EDWARD KULICK, ASSESSING UNEXPECTED DIFFERENTIAL ITEM PERFORMANCE OF FEMALE CANDIDATES ON SAT AND TSWE FORMS ADMINISTERED IN DECEMBER 1977: AN APPLICATION OF THE STANDARDIZATION APPROACH (1983), *available at* www.ets.org/research/researcher/RR-83-09.html; Neil J. Dorans & Edward Kulick, *Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test*, 23 J. EDUC. MEASUREMENT 355, 355–56 (1986).

[24] *See* Freedle, *Correcting the SAT's*, *supra* note 6, at 3; Freedle, *A Review of Dorans*, *supra* note 6, at 75, 77.

as well), should have the result of increasing their representation in many of the elite universities.

Table 2 presents a few examples of DIF bias as it occurred in earlier versions of the SAT.

Table 2

EXAMPLES OF HOW ANALOGY ITEMS FROM DISCLOSED SAT TESTS FROM THE 1980s UNFAIRLY DISCRIMINATED AGAINST AFFRICAN-AMERICANS*

2 HARD ANALOGY ITEMS:	2 EASY ANALOGY ITEMS:
Surreptitious: stealth a) clandestine: secrecy b) subversive: unity c) omnipresent: generosity d) verbose: enunciation e) opulent: simplicity [a is correct]	Crest: wave a) trunk: tree b) shore: lake c) hub: wheel d) base: triangle e) peak: mountain [e is correct]
Sycophant: flattery a) impostor: deference b) embezzler: insolence c) bandit: hypocrisy d) swindler: fraudulence e) advocate: defamation [d is correct]	Bark: tree a) skin: fruit b) dew: grass c) seed: flower d) peak: hill e) wake: boat [a is correct]

*Note on Analogy items: Semantic analyses, *see* Freedle, *Correcting the SAT's*, *supra* note 6; Freedle, *A Review of Dorans*, *supra* note 6; Freedle & Kostin, *Predicting Black and White*, *supra* note 22; Freedle, Kostin & Schwartz, *supra* note 22; FREEDLE, KOSTIN & SCHWARTZ, *supra* not 22, showed that if HARD analogy items contained a social-personality type content and used rare vocabulary concepts, African-Americans performed significantly BETTER than matched-ability White students. However, if the same matched-ability students were compared on the EASY analogy items, then African-American students performed significantly WORSE than White students when very easy (common, everyday) vocabulary concepts were used. It is hypothesized that ethnic groups differ quite widely in the various semantic senses that common words can assume. *See* Freedle, *supra* note 6. Unfortunately, there is no study that explores the various semantic senses frequently used by each ethnic group, so the hypothesis is hard to verify. A similar pattern of item bias was found for sentence-completion items. *See* Freedle, *supra* note 6.

At this point, I need to alert the careful reader that there are at least two ways in which a test can be racially biased. The first way a test can be biased has already been discussed vis-à-vis the Fagan and Holland approach—it involves assessing only the mean difference in performance for any two populations.[25] The Fagan and Holland result, you will remember, focused upon the *total scores* achieved by each racial group.[26] Performance on each individual test item is not of special interest in this first method of evaluating test bias. By contrast, the second way a test can be biased comes from an examination of *each test item*, taken one at a time. It should be clear that many standardized tests can be biased in both ways: there is a mean difference, say, between two ethnic groups, and furthermore there is often systematic individual item bias.[27]

It seems reasonable to expect that if a test has both types of bias as I maintain existed, especially for the early SAT tests administered in the 1980s, and if one removes most of the item bias from more recent versions of the test, then one would expect the mean performance of the two ethnic groups (African-Americans and Whites) to converge in the manner described in my 2003 paper.[28] Well, this is not what happens and the reason it does not happen brings us to another ugly little chapter in standardized testing: something I call the Rosner Effect.

Before we explore the Rosner Effect, let me first draw a conclusion regarding item bias effects as embodied in the DIF technique. The good news is—for standardized tests like the SAT, and, presumably this applies as well for the LSAT—one *can* diminish the magnitude of the DIF bias for individual test items even though a small residual effect still lingers.[29] But the really bad news, as I have suggested, is the Rosner Effect. The Rosner Effect totally negates any reduction in item DIF bias in the subsequent development of new "parallel" test forms. That is, you can eliminate one form of bias fairly easily (the individual item bias effects). This is accomplished by avoiding certain item

---

[25] *See* Fagan & Holland, *supra* note 1, at 362, 382 (analyzing Jensen's biased *default hypothesis*).

[26] *See id.* at 368–76.

[27] Freedle, *Correcting the SAT's*, *supra* note 6, at 2.

[28] *See id.* at 1, 9–11.

[29] *See* Freedle, *A Review of Dorans*, *supra* note 6, at 73.

formats and certain contents.[30] I thought that applying these corrective measures to individual items would move the two ethnic populations of students closer together. However, I never accounted for the existence of the Rosner Effect. The Rosner Effect guarantees that the first form of test bias will be perpetuated, in the sense that the same mean difference is maintained between any two ethnic groups when new test forms are developed. Let's deal with it now.

A. The Jay Rosner Effect

The name of Jay Rosner should be familiar to readers of this journal. Rosner's argument on behalf of Affirmative Action in the recent Supreme Court affirmative action decisions showed that the usual statistical procedures that are in place for developing new test forms of equivalent difficulty for any standardized test almost guarantee the perpetuation of ethnic bias in newly developed test forms.[31] Rosner explained that when a new test form is constructed by selecting items from a large number of potential test items, test assemblers routinely select *only* items that favor Whites over African-Americans.[32] This is done even though one can find in the batch of potential items many examples wherein the African-Americans outperform the White students.

Rosner explained that the reason mean differences among the races persist in test form after test form is that the statistical procedures in place for building equivalent "parallel" forms from one year to the next are specifically devoted to perpetuating these differences.[33] Statisticians routinely justify this selection process by indicating that one needs to build a new test that is as reliable, valid, and replicable as previous test forms. While this may sound good on its face, it ignores the Fagan-Holland test results. That is to say, because there is no intrinsic difference in verbal ability across races when students are given an equal

---

[30] *See* Freedle & Kostin, *Predicting Black and White*, *supra* note 22, at 425–27, 436–37 (including examples such as avoiding part-whole relationships when writing analogy items or by substituting a more philosophical approach to science content instead of a direct excerpt approach).

[31] *See generally* Gratz v. Bollinger, 539 U.S. 244 (2003); Grutter v. Bollinger, 539 U.S. 306 (2003).

[32] Rosner, *supra* note 18, at 24.

[33] *Id.*

opportunity to learn and reason about the test materials, perpetuating old racial biases in existing tests simply because earlier tests started out racially biased is not a sufficiently persuasive rationale for continuing this abuse. So, I would caution the statisticians to reconsider what they are doing. They should become more culturally conscious and let it inform their statistics. From another perspective, I would say that most psychometricians, statisticians who work on constructing tests, are probably well-intentioned, but the traditional test procedures that are in place compromise their good intentions. The only real solution for the ethnically negative Rosner Effect, in my opinion, is to replace currently biased tests with the unbiased Fagan-Holland test when assessing verbal aptitude. Other tests that involve the assessment of achievement such as the Advanced Placement tests of the College Board, the subject area tests of the SAT II, the SAT math test, and so on, still need to deal seriously and directly with the corrosive Rosner Effect. As we have just seen, my Revised-Score method for correcting some of the ethnic bias at the individual item level, even for achievement tests,[34] will be seriously eroded by the application of the Rosner Effect in the development of future test forms for such tests.

B. Conclusion Regarding the Rosner Effect

Admissions officers should consider the Fagan-Holland test as a reliable and unbiased measure of verbal ability. A much weaker conclusion is the following: although it sounds dishonest to suggest it, suppose all students either leave blank all questions about their ethnicity or decide to say they are "White" when they sit down to take a standardized test (after all, all races are intermixed to some degree and one can probably legitimately claim to be a member of each and every race to some degree). In this way, it would be much more difficult for the statisticians to select new items for any new test form that intentionally perpetuates mean ethnic differences that exactly mirror the old ethnic differences of past test forms. "Will the desired end (greater racial fairness) justify the means (withholding racial information)?" It is a moot point, but I would not attempt to answer it until I heard the answer given by test

[34] *See* Freedle, *Correcting the SAT's*, *supra* note 6, at 28–29 (regarding Advanced Placement test bias and SAT-math bias).

organizations after they ask themselves a parallel question: "Does the desired end (continual flow of income) justify the means they currently employ to get it (constructing racially biased tests)?"

III. A NEW STATISTICAL MODEL FOR ASSESSING THE LEVEL OF GUESSING: INTRODUCTORY STATEMENT

Even though I prefer the Fagan-Holland test of verbal aptitude to any other standardized test of verbal aptitude, and even though their test could easily be implemented as a well-controlled test via a computer terminal, realistically it seems unlikely that educational institutions will rush in to consider adding this test to their list of requirements. Unfortunately, I believe these institutions—like the testing establishment—are also wedded to traditional ways of selecting students, even if these methods are unwieldy, distressing, and ultimately biased. I recognize that the Fagan and Holland type test can only go so far in correcting the persistent verbal bias problem in testing. For one thing, other standardized tests, as mentioned above, are still needed to assess levels of performance on achievement tests (such as the Advanced Placement tests, the subject-area tests of the SAT II, the math section of the SAT, etc.). Facing this broader reality means that *one should not give up solving other persistent problems in traditional test theory*, especially if such advancements might further mitigate the wide divide among ethnic groups. A problem that definitely affects ethnic evaluation, especially for lower-scoring students, on such tests such as the verbal SAT is the matter of *guessing* and how to properly measure it.

In this section, I first will present several of the details of the theory in a manner that should largely be accessible to the layperson. Then I will illustrate how this new guessing theory has practical applications regarding the rates at which African-American and White students are guessing on standardized tests. But first, the theory.

A. Solving the Problem of Guessing On Tests: A Solution in Plain View for Fifty Years

It is necessary to write about the issue of guessing when it comes to standardized testing, including tests such as the LSAT, because researchers at the College Board and the Educational

Testing Service have expressed the view that minorities are "lucky guessers" when it comes to how they respond to hard test items.[35] That is, the notion of "lucky guessing" is at the core of their early criticisms of my studies that showed that minorities at all score levels are systematically getting more hard items correct when compared with matched-ability Whites.[36] These testing "experts" have explained that the only reason minorities perform so much better than they are "supposed" to do on hard verbal and hard math items is that they are "lucky."[37] I intend to show in this section that their idea of "lucky guessing" is hopelessly outmoded and theoretically misguided.

The purpose of this section, therefore, is to introduce a very simple mathematical solution to the 50-year old problem of guessing. The solution actually applies to all multiple-choice tests, regardless of whether minorities are involved or not. I first present a numerical illustration of the new guessing theory using White female[38] responses to SAT analogy items[39] administered in the mid-1990s. I later apply the results of this new theory to estimate the actual minority and White guessing levels as they apply to 12 analogy items taken from an SAT exam administered in the 1990s. As we shall see, the College Board's "generous" estimate of who is guessing and who is not is quite wrongheaded.

B. An Illustration that Leads to a New Empirical Estimation of Guessing on Standardized Multiple-Choice Tests

Most psychometricians[40] suggest that when twenty percent or less of a group of examinees have correctly answered a particular test item (for example, an item with five response options), then ALL their responses to such an item must be considered to have been randomly guessed.[41] I propose that this

---

[35] Freedle, *A Review of Dorans*, *supra* note 6, at 75–77, 79.

[36] *See id.* at 75–77; *see also* Freedle, *Correcting the SAT's*, *supra* note 6, at 3.

[37] Freedle, *A Review of Dorans*, *supra* note 6, at 75–77, 79.

[38] I could have used any group of students; however, to avoid further pointless criticisms, I have chosen to apply the theory initially to White students.

[39] Taken from SAT form QSA-09.

[40] A psychometrician is "a person (as a clinical psychologist) who is skilled in the administration and interpretation of objective psychological tests (as of intelligence or of personality)." WEBSTER'S THIRD NEW INTERNATIONAL DICTIONARY 1833 (1993).

[41] Ronald K. Hambleton, *Principles and Selected Application of Item Response Theory*, *in* EDUCATIONAL MEASUREMENT (Robert L. Linn ed., 3d ed. 1993).

conventional approach to guessing is wrong.

(1) Below, I will first illustrate what truly random data looks like followed immediately by what the responses to a typically hard item actually look like.

(2) A little later, I will describe how to determine the *actual level of guessing* that occurs for all items, regardless of their difficulty level and regardless of the ability level (or the ethnic background) of the students responding to these items. This illustrates how broadly applicable this new procedure is.

(3) Finally, by way of illustration, I will focus upon the responses of White and matched-ability African-American students who have received low SAT scores (from 200 to 260) on twelve analogy items taken from a particular SAT form.[42] I later use this data for twelve analogies in order to illustrate how to apply this new guessing theory so that it informs us about racial differences and the different levels with which each ethnic group uses guessing when responding to all multiple-choice test items. Again, I emphasize, this new guessing theory applies to all multiple-choice items and to students of all ability levels, regardless of ethnic background.

Under point 1, I want to illustrate what a *truly* random procedure would look like. Suppose you have a test item with five response options consisting of options a, b, c, d, and e. To keep things simple, let's suppose that the first option is always the correct answer.[43]

If you have about 800 students responding in a truly random manner, you would get something like the following percentages of choices for the five response options: 18%, 21%, 22%, 20%, 19%. Notice that each percentage remains very close to the 20% chance level hypothesized for a truly random process; in the example you get small departures amounting to at most about 2% points above or below the pure theoretical value of 20%. The problem is, when you look at real data you will find the pattern of choices looking something more like the following: 15%, 19%, 12%, 16%, 38%. You will notice that one of the wrong options (here, the fifth option) is selected by a very large percentage (38%) of the students. A statistical test shows that such an item

---

[42] SAT form QSA-09.

[43] The following example is adapted from one that I presented in an earlier paper. *See* Freedle, *A Review of Dorans*, *supra* note 6, at 75–77.

wildly departs from a statistical test for randomness. Yet, most test "experts" classify such an item as falling within the category of pure "guessing" (with a *full 100% of the students presumed to have randomly guessed* their answers) simply because the correct option is chosen by close to (or less than) 20% of the students. Indeed, to make matters even more embarrassing for such test experts, an acknowledged authority in test theory has said that guessing

> [i]s included in the model to account for item response data from low-ability examinees . . . where . . . guessing is a factor in test performance. It is now common to refer to the [guessing] parameter . . . as the *pseudochance level* or pseudoguessing parameter [which typically] takes on a value that is smaller than the value that would result if examinees of low ability were to randomly guess the item.[44]

In other words, "chance" is not really chance—it is pseudochance. Well, that is not very enlightening, is it? Indeed, I have examined hundreds of test items for low (and even middle) ability students of all ethnic groups, and it is clear that in every case none of the data fit the classic definition of truly random data. So, it should be clear that *all* students are not simply guessing when the correct answer that is selected just happens to have been selected by around 20% (or fewer) of the students. So, what is the alternative? What is a better estimate of how many students really have been guessing? To answer this let me present you with a little more data.

The data I present below is taken from SAT test form QSA-09 (presented in the mid-1990s) for White females with English as their best language. There is no special reason for selecting this particular sample of students; it is simply a sample of data that I had already analyzed. Later, as suggested above, ethnic data will be presented to illustrate a very specific criticism raised by Educational Testing Service researchers against my use of the "guessing" concept.

The data in Table 3 are all for one particular analogy item, item 16, taken from test section 1 of test form QSA-09. This item was selected because the lowest scoring individuals with SAT scores of 200 and 300 selected the correct option (for this example it is the *third option* that is correct) at substantially less than the

---

[44] Hambleton, *supra* note 41, at 147, 155.

20% correct level.[45]

Table 3

NEW GUESSING MODEL FOR SAT ANALOGY ITEM 16 (SECTION 1) FROM THE SAT FORM QSA09 (MID-1990s)

SAT Score	Percent Selecting Each of Five Options					Percent Blanks	[Percent Blanks / 5]	[M+]= [(Minimum + (Blanks) / 5)]	New Guess: 5*[M+]	Old Guess Model
	1st	2nd	3rd	4th	5th					
200	15.7	25.1	(07.6)*	14.1	27.6	09.9	[09.9 / 5 = 2.0]	07.6 + 2.0 = 09.6	48.0	100.0
300	21.5	18.9	(16.6)	06.7*	17.3	19.0	[19.0 / 5 = 3.8]	06.7 + 3.8 = 10.5	52.5	100.0
400	16.9	13.1	(41.5)	03.5*	12.1	12.9	[12.9 / 5 = 2.6]	03.5 + 2.6 = 06.1	30.5	0
500	10.4	06.8	(69.9)	01.1*	06.9	04.9	[04.9 / 5 = 1.0]	01.1 + 1.0 = 02.1	10.5	0
600	04.8	03.1	(87.2)	00.9*	02.4	01.6	[01.6 / 5 = 0.3]	00.9 + 0.3 = 01.2	06.0	0
700	01.8	00.2	(97.0)	00.0*	00.8	00.2	[00.2 / 5 = 0.0]	00.0 + 0.0 = 00.0	00.0	0

* Indicates the smallest entry among the five response options at each of the 6 ability levels. [Note, it is possible for the smallest entry to coincide with the correct option. Also it is possible that different options will prove to be the smallest entry as one moves up the ability level. For this particular item, we see that the smallest entry is the 4th one for all ability levels except for those with an SAT of 200.] The fact that most of the students reject the 4th option as obviously incorrect shows that responses are not random even for students of lower ability. We also see, comparing the last two columns, that there is considerable difference in measuring how much guessing occurs for each ability group in the new guessing model as compared with the old model.

Let us look even more closely at some of the numbers listed in Table 3. The people who earned an SAT score of, say, 300 selected the first option 21.5% of the time, the second option was selected 18.9% of the time, the third option (the correct option) was selected only 16.6% of the time, the fourth option was selected 6.7%, and so on. You will also notice that 19.0% of these students either omitted or never reached this particular option; this number (19.0) is entered under the "Percent Blanks" column and we will be using this number later to refine our new model of guessing. My new guessing model says that for every test item and for every ability level one can immediately get a rough estimate of how much guessing has occurred by finding that response option with the smallest entry. For the people with SAT scores of 300, the smallest entry just happens to be 6.7 (which was the fourth response option). The particular version of the model that I am presenting to you now (there are other versions) says that, for this particular item, every response option is "contaminated" with at least 6.7% of guessing for this particular ability group. To get rid of the contamination you

[45] People with an SAT score of 200 selected the correct option only 7.6% of the time, while those scoring 300 selected the correct option only 16.6% of the time—both lower than 20% which the "experts" consider an especially noteworthy value.

subtract the 6.7 from each of the five response options. This leaves you with the following values: 21.5–6.7 = 14.8; 18.9–6.7 = 12.2; 16.6–6.7 = 9.9; 6.7–6.7 = 0.0; 17.3–6.7 = 10.6. Or more simply: 14.8, 12.2, 9.9, 0.0, 10.6.[46] So, if my model proves to be consistent with further theoretical evaluation (see below), you can see how the answer to this 50-year old puzzle in test theory concerning how to measure guessing more accurately was in full view for everyone to see, but no one has seen it until now. The "experts" were too obsessed with the magical number of 20% or less for just the correct option to examine the theoretical import of the option chosen by the *least* number of students. There is here theoretical power in being amongst the least of the entries.

You will also notice that Table 3 shows us that even the highest ability groups are engaged in some level of guessing, no matter how small. For example, 6% of the students who earned an SAT of 600 were engaged in guessing—this is shown in the column labeled "New Guess: 5*[M+]." None of the people with SAT scores of 700 however were guessing, as Table 3 indicates. The important point here is that we are able to estimate guessing for each and every ability level! That was not possible with earlier models of guessing.

How can one determine whether what I have just presented is an accurate estimate of how much guessing really is occurring? First, I need to present a very simple picture of how statisticians try to boil all the data down into a manageable set of numbers (called model parameters). Then, I will show you how to evaluate whether the guessing procedure described above really makes sense or not.

C. *A Simple Picture of Item Response Theory ("IRT") for Either Two Parameters or Three Parameters*

When the IRT model uses *three parameters*, each parameter describes different ways in which the whole group of students has responded to a test item. The first parameter is used to describe the overall difficulty of an item, the second parameter (called item discrimination) describes whether there is a sharp

[46] Later, I will present more details about what to do about the percent blanks when you are correcting for the guessing contamination. For this example, what I have just presented actually represents a correct result, but the full exposition of how we get to this point is a little too complicated for our present purposes of providing a clear step-by-step exposition.

(or a gradual) decrease in item difficulty as the ability of students gets higher and higher, and the third parameter is used to describe whether many of especially the low ability students are guessing. When the IRT model uses just *two parameters*, one measures just the item difficulty and the second measures just the item discrimination (as described above). No guessing is assumed to occur when one fits just a two-parameter model to the data.

For ease of reference, let's agree to call the three-parameter model just IRT-3 and the two-parameter model IRT-2. Basically, the IRT-3 and IRT-2 models try to fit all the percent correct responses of all the ability groups with a single mathematical curve called a logistic. Typically, IRT-3 is needed to fit all the hardest items because more students are assumed to be guessing when an item is quite difficult. IRT-2 might be used to fit some of the easiest items where guessing is probably minimal; but again, in practice, an IRT-3 model is probably used to fit even these easy items. After all, if it turns out that the students have not been guessing, then the guessing parameter will be estimated to be zero. All of what I have been saying so far, as IRT models go, is just standard test theory.[47]

Suppose we examine a difficult item and suspect that many of the students have been guessing; so, we will certainly try to fit a three-parameter IRT-3 model to such data. Now, it should be obvious to the reader that if there is a simple way to measure how much people are actually guessing (by just examining the raw data for each item), one should be able to take this simple estimate and subtract it from the observed percent correct. If we do that, we automatically will convert the data from an IRT-3 problem into an IRT-2 problem. By first subtracting out the guessing component, we will therefore have saved ourselves some work in finding a quick solution to the parameter estimation problem because we no longer have to separately estimate the third (guessing) parameter—it has been eliminated by the subtraction process. All we have left to estimate is the item difficulty parameter and the other parameter (the discrimination parameter).

I will now show you, with real data, what this looks like. For

---

[47] *See* Hambleton, *supra* note 41, at 155.

analogy item 16,[48] for White female students, the raw percent correct responses for each of six ability groups (with SAT scores of 200, 300, 400, 500, 600 & 700, respectively) was:

7.6 16.6 41.5 69.9 87.2 97.0 (Step 1)

These values mean that of those with an SAT score of 200, 7.6% of them selected the correct option, while at the other extreme those scoring 700 on the SAT selected the correct option 97% of the time.

We are going to correct these raw percent correct responses by subtracting out the smallest entry for each of these six ability groups. The smallest entries (what I call here the "raw" guessing estimates) for each of the six ability groups are as follows:

7.6 6.7 3.5 1.1 0.9 0.0 (Step 2)

Notice that for Step 2, as we move from the lowest ability students to the highest ability students, the estimate of how much true guessing was going on gets generally smaller and smaller so that at SAT 700 none of these students (i.e., 0%) were guessing. We now subtract each guessing estimate in turn from each of the original raw percents correct responses and this gives us:

0.0 9.9 38.0 68.8 86.1 97.0 (Step 3)

I propose that because we have extracted all the guessing responses from these data, that what we have left in Step 3 should be well-fitted by just a two-parameter (IRT-2) model. If in fact this turns out to be well-fitted by a two-parameter model, it is partial evidence in favor of our new guessing model. Let us further say that we get a best-fitting estimated value of -.50 for the first parameter and a best-fitting value of 1.40 for the second parameter. These two parameter values (-.50, 1.40) are in fact the best-fitting IRT-2 values for the data presented in Step 3. Keep these values in mind as I develop the rest of the argument for the new guessing model immediately below.

Now, before we move forward with the demonstration of the

[48] Taken from SAT form QSA-09, section 1.

new guessing model, I want to tell you that I purposely avoided mentioning a complicating factor in obtaining this last set of values (i.e., the 0.0, 9.9 . . . 97.0 entries listed in Step 3). I need to mention it here in order to avoid confusing the reader in what follows. Remember that all students at each score level typically have a small percentage of responses that are left blank—students either purposely omitted the item (these are called "O" responses) or they never got to consider the item (these are called "NR" responses for "not reached"); these NR responses might happen with the hardest items that occur at the end of many test sections where some students simply have not worked fast enough to get to the final items. There are several ways to deal with these blank responses. The simplest assumption is that if the students had been forced to fill in these blanks, they would have picked the available five options at random (this is called Model A); that is, they would have been strictly guessing at random. So now you say: "Well if these represent additional guessing, why did you not add these values into the smallest entries which would slightly inflate the estimate of what the true guessing level really should be?" You would be right in raising such an objection. The only reason I did not include it early on is that it *would not have changed the final results* presented in Step 3 above. And trying to present this complication so early in the demonstration would probably have added more confusion than light. But now I must deal with the issue before we can move forward.

Let me show you why it would not have changed the values in Step 3. There are five response options, each with its percentage of being selected plus a percentage of blanks—let us say there are 10% blanks. All of these values sum to 100%. Suppose there are 30% raw *correct* responses and the *smallest* of the five options is 5% (which would be our initial estimate of the amount of guessing for this group of students). Remember, above I said let's just subtract 5 from 30 to get a revised estimate (25%) of what the true correct responses should have been that is totally free of guessing. Okay, but now we want to know in more detail what to do with the, say, 10% blank responses. We have already said that we assume that if students had been forced to fill in these blanks they would have randomly selected among the five response options. So, 10%/5 = 2%. That means that we should add about 2% to *each* of the five response options and that

gets rid of worrying about what to do about the missing data (the blanks). Okay. That means that there are now 32% (30+2) apparently correct responses and 7% (5+2) apparent guesses. If we subtract 7% from 32% we get 25% which is *exactly* the amount that we started with. So nothing is numerically changed by this digression about what to do with the missing data. That is why I early on avoided going into detail concerning it. But in the next section we are going to have to deal with the missing (blank) data more explicitly. Now I will continue with the demonstration of the new guessing model.[49]

The original raw data for percent correct responses (given above in Step 1) should be well-fitted only by a three-parameter (IRT-3) model because it still contains the guessing responses. Actually, we need to do one last thing to these raw data (the original percent correct response data) before we can move forward and fit these data with a mathematical equation. As suggested above, we need to deal with the question of what to do with the "blank" responses. For the six ability groups, respectively, the percentage of blank responses was:

9.9 19.0 12.9 4.9 1.6 0.2 (Step 4)

We are going to continue to assume for one version of our guessing model (called Model A) that all of these blank responses would have represented an additional source of pure guessing had the students been forced to select one of the five response options. So we divide each entry by a value of 5.0 to find out how much additional guessing would have been present in each of the five response options—this additional amount represents an artificial "inflated" value that needs to be added to the raw percent correct responses. If we divide each of the entries in Step 4 by the value 5.0, we get (after rounding off):

2.0 3.8 2.6 1.0 0.3 0.0 (Step 5)

Now go back to the original raw correct data which was:

---

[49] The reader, however, should carefully note that there are *other* models beside Model A wherein a more complicated argument would necessarily ensue concerning what to do about the missing data. Fortunately, Model A, as presented, conveniently bypasses several of these complications.

7.6 16.6 41.5 69.9 87.2 97.0 (Step 1, repeated for convenience)

Add the values from Step 5 to the values of Step 1, and we get:

9.6 20.4 44.1 70.9 87.5 97.0 (Step 6)

Step 6 represents the correct responses fully contaminated by guessing. Therefore, the data in Step 6 should be well-fitted by a three-parameter model (IRT-3). Remember that we found a little earlier that the first parameter (the difficulty parameter) was estimated to be -.50 for the IRT-2 model, and its second parameter was estimated to be 1.40—this was for the data presented in Step 3. If the only difference between the data in Step 6 (fully contaminated by guessing at each ability level) and the data in Step 3 (uncontaminated by guessing at each ability level) turns out to be the presence versus the absence of appropriated measured guessing levels, then we should expect to find that the best fitting first and second parameters for the data in Step 6 will be very similar (give or take a few points due to measurement error) to those values already found to the IRT-2 model (namely, the -.50 value for the first parameter and the 1.40 value for the second parameter). And that is what happens. The best-fitting IRT-3 parameter values for the data in Step 6 yields -.50 for the first (difficulty) parameter, 1.40 for the best-fitting second parameter (the discrimination parameter), and .10 for the guessing parameter. Such close agreement between the first two parameters (for IRT-2 versus IRT-3) does not always happen. But of the many sets of test items that I have already fitted (which includes reading comprehension items, analogy items, and some math items), the agreement between IRT-2 and IRT-3 parameters is generally quite close. For readers who are curious, if the values in Step 6 above (9.6, 20.4, 44.1, 70.9, 87.5, 97.0) are called the "observed" values, then the predicted values (using the three parameters -.50, 1.40 and .10) are: 12.0, 19.0, 39.0, 70.0, 90.0, 97.0. Most of the paired values are in excellent agreement; the most deviant pair of current values (44.1 from the "observed" list versus 39.0 from the predicted list) looks like it might be improved somewhat perhaps by using other methods of estimating parameters—the method I used for best-fitting

parameters was the smallest absolute deviation between the set of observed and predicted values. But we should be quite pleased, nonetheless, with the fact that for this particular example, there is such a close match between the first and second parameters of both IRT-2 and IRT-3 models.

Let us continue with this issue of fitting the "observed" data with some "predicted" values. I want to return briefly to the values listed in Step 3 above (0.0, 9.9, 38.0, 68.8, 86.1, 97.0) which, you will remember, were the "observed" values for the IRT-2 model with all the guessing behavior removed. I said earlier that the best-fitting parameter values for these data were -.50 and 1.40. When these values are in fact inserted into the equation for the IRT-2 model, the "predicted" values that emerge from it are, respectively: 2.0, 10.0, 33.0, 66.0, 89.0, 97.0. The largest departure between "observed" and "predicted" values is again for the third set of entries (namely, 38.0 versus 33.0), but all the other paired values appear to be adequately close. Again, it is possible that other methods of finding the best-fitting parameter values might lead to an even closer fit between "observed" and "predicted" values. That remains for future explorations.

Now, before I apply this new guessing model to contrast the guessing behaviors of African-American and matched-ability White students, let us examine a few other points about guessing.

Let us compute how much overall guessing was occurring on this analogy item[50] for White females. To begin to find the overall amount of guessing we first locate again the minimum response percent for each ability level—here from SAT 200 to SAT 700, respectively:

7.6 6.7 3.5 1.1 0.9 0.0 (Step 2, repeated for convenience)

And to these values we add one-fifth the percent of items left blank which is, respectively:

2.0 3.8 2.6 1.0 0.3 0.0 (Step 7)

---

[50] SAT form QSA-09, item 16, section 1.

The result of this addition is:

9.6 10.5 6.1 2.1 1.2 0.0 (Step 8)

This means that, say, for people with SAT scores of 200, it is estimated that 9.6% guessing has occurred based on the smallest entry among five response options. But the value of 9.6% actually applies to *each* of the five options! That is, this guessing estimate actually contaminates each and every one of the five options. Therefore, to estimate the overall guessing rate that has occurred for people getting SAT scores of 200, we multiply each value in Step 8 by a factor of five to get (for the first entry) 9.6*5 = 48%. This says that of all the people who got an SAT score of 200, nearly half of them (48%) were just guessing for this particular analogy item. But this figure of 48% of the students were guessing is quite different from the figure of 100% that conventional theory says were guessing. Remember, because fewer than 20% of the students were getting this item correct, conventional theory concludes that *everyone* must have been guessing! *It is clear that the old theory and my new guessing theory lead to very different conclusions*! And I further submit that, in light of this new theory, the old theory is woefully inadequate in how it has explained and used the phenomenon of guessing. So, to see how much guessing occurred for each of the six ability groups for this analogy item, we multiply each entry from Step 8 by 5.0 and we get:

48.0 52.5 30.5 10.5 6.0 0.0 (Step 9)

Conventional theory says that the first two entries of Step 9 should both be 100% of guessing, while the last four entries should be zero (if one assumes a "threshold" approach to guessing). It is obvious again that my new guessing theory leads to a more nuanced approach with a smoother gradient of guessing values than the abrupt "threshold" effect that conventional theory claims is correct.[51]

---

[51] The old "threshold" model of guessing merges with my new definition of guessing only under the following circumstances. Regardless of which option is chosen, when the minimum option (out of n options) is approximately of the size 1/n (expressed as a proportion), then the threshold model and my new guessing model yield the same prediction. That is, nearly all students can be truly assumed to have

So far, the "observed" and "predicted" set of values appear to be very similar. We can use some other statistical procedures to describe the full set of 13 fitted analogy items to judge the overall similarity between "observed" and "predicted" values. When we examine all 13 analogy items from test form QSA-09 for White male students with English as their best language, we generate for each analogy item six predicted values and six observed values (for each of the 6 ability levels of 200, 300, 400, 500, 600 and 700, respectively). Since there are 13 such items there are a total of 78 (13x 6 = 78) paired sets of values. We first correlate the observed and predicted values for just the IRT-2 model.

This yields a very high value of 0.9966 (where a perfect agreement between observed and predicted would have produced a value of 1.0000). This is highly significant ($p < .001$). The mean value for all the predicted entries was 53.42 while the mean for the observed values was 53.23; this shows us that there is a very close similarity between observed and predicted values for IRT-2 data.

On the other hand, for observed and predicted IRT-3 values (for which there are again 78 pairs of entries) the overall correlation, while high, is not as large—the actual value was .9630 ($p < .001$). If we examine the mean predicted values for IRT-3 it was 59.24 while the mean observed values for IRT-3 was 57.63. We see that there is a tendency for the predicted values to be in excess by the amount of 1.61%.

Comparing IRT-2 with IRT-3 results we see that there is, at least for these 13 analogy items, a much better fit obtained in describing the data once all guessing responses have been omitted (the IRT-2 model) than when all guessing responses have been added to the raw correct responses (the IRT-3 model)! In other versions of this new guessing model, I explain why this tends to occur.[52]

D. General Evaluation of the New Guessing Theory

Table 4 lists the agreement between the *a* and *b* parameters

---

been guessing. For example, for n=5 (i.e., for five options), when any option is approximately 20% and is at the same time the minimum entry out of five options, then it can be concluded that most of the students have been guessing.

[52] Roy Freedle, New Models for the Direct Estimation of Guessing at Each Ability Level for Standardized Multiple-Choice Tests (Jan. 5, 2005) (unpublished manuscript, on file with author).

for all 13 analogy items taken from SAT form QSA-09. The data, without loss of generality in demonstrating the new method, was restricted to White male students with English as their best language. One can see that, in general, my new guessing theory yields a *very close agreement*, between the a parameter estimated for the IRT-2 model and the a parameter estimated for the IRT-3 model: yielding a mean absolute difference of .09 in the a values. As Table 4 also indicates, the mean absolute difference of .04 occurs in estimating the b parameter values. These close agreements provide strong empirical support for the validity of my guessing theory.

Table 4

THE BEST FITTING PARAMETER VALUES FOR THE 2-PARAMETER IRT MODEL COMPARED WITH THE 3-PARAMETER IRT MODEL FOR 13 SAT ANALOGY ITEMS: AN EVALUATION OF A PREDICTION MADE BY THE NEW GUESSING THEORY

	Model A*		Model A		
Analogy	**IRT-2**		**IRT-3**		
Item	***b***	***a***	***b***	***a***	***c***
1-11	-3.00	0.90	-3.00	0.90	0.00
1-12	-3.00	1.90	-3.00	1.50	0 .05
1-13	-2.80	1.20	-2.90	1.20	0.05
1-14	-2.20	0.60	-2.00	0.50	0.10
1-15	-2.00	0.70	-2.00	0.70	0.05
1-16	-0.80	1.40	-0.80	1.30	0.10
1-17	-0.40	1.30	-0.40	1.30	0.15
1-18	-0.20	0.90	-0.20	0.90	0.10
1-19	-0.40	0.90	-0.50	1.80	0.15
1-20	0.40	1.00	0.40	0.90	0.15
1-21	0.60	1.20	0.60	1.20	0.15
1-22	1.80	0.60	1.90	0.50	0.15
1-23	2.00	1.00	2.00	1.00	0.15
Mean (n=13)	-0.77	1.09	-0.76	1.05	
S.D. (n=13)	1.73	0.38	1.73	0.38	
Mean Absolute deviation for the two *b* parameters: .50/13 = .04					
Mean Absolute deviation for the two *a* parameters: 1.10/13 = .09					

* *b* = item difficulty parameter, *a* = item discrimination parameter, *c* = guessing (for IRT-3 only). The best fitting parameters *b* and *a* are indicated for just Model A for IRT-2, and, the best fitting parameters *b*, *a*, and *c* for IRT-3 are also presented for just Model A. The criterion for best fitting solution was the least cumulative absolute difference. Analogy items are from SAT disclosed form QSA-09. Analogy item 1-17, while presented in this table, is removed from Table 5 in this report in order to simplify the contrast between easier analogies (n=6) and harder (n=6) analogies. The *new guessing theory* asserts that the *b* parameter of IRT-2 should be identical (or similar in magnitude to) the *b* parameter of IRT-3. The same applies to the *a* parameter in comparing across IRT-2 and IRT-3 models.

Because it is instructive, I will briefly mention another

model (called Model B) that was developed to account for the Blank Responses (that is, all the Omitted and Not-Reached items). Model B evaluated the possibility that all Blank Responses should be proportionately distributed among the five available options—that is, Model B assumed that if the students who omitted items or failed to reach items were forced to fill in the blanks they would have distributed their responses, not randomly, but proportionate to those responses already contained among the five options. The stability of the parameter values for *a* and *b* did not provide as good a fit to the data as our earlier Model A provided. This can be seen in the fact that the mean absolute deviation of parameter *b* (for Model B) was .15, whereas for Model A, the mean value was only .04. Similarly, the mean absolute deviation of parameter *a* (for Model B) was .11, whereas for Model A the mean value was .09. Therefore, regarding both parameters, Model A clearly provides a more stable fit when contrasting IRT-2 with IRT-3 parameter estimates. This nicely illustrates how model fitting helps one narrow down the possibilities of what students are actually doing when they take tests.

In a forthcoming paper, I describe two other possible models for fitting the analogy data,[53] but this would take us too far afield to describe these new models. Suffice it to say that the large correlations showing the closeness of the predicted and observed values (i.e., the 78 paired values cited earlier) along with the close similarity of the parameter values themselves (for Model A) provide encouraging support for my new theory of guessing.

E. Practical Application of the New Guessing Model to Low Ability Students of Two Races

Now I present one practical outcome of this new guessing model. Table 5 shows what happens to performance (*for the lowest scoring students with SAT verbal scores of 200 to 260*) on the six easiest and the six hardest analogy items taken from SAT test form QSA-09 administered in the mid 1990s. Contrary to what researchers at ETS have claimed,[54] African-American

---

[53] *Id.*

[54] Brent Bridgeman & Nancy Burton, Does Scoring Only the Hard Questions on the SAT Make It Fairer?, Address at the 2005 Annual Meeting of the American Educational Research Association (Apr. 12, 2005).

students who get very low SAT scores in fact are performing better than so-called matched-ability White students do on the hardest analogy items (which use rare vocabulary words) and worse on the easiest analogy items (which employ very common vocabulary words). Instead, Bridgeman and Burton suggest that the responses of low-scoring students are hopelessly mired in purely random responding.[55] The data clearly contradicts this.

Table 5

THE APPLICATION OF A NEW MODEL OF GUESSING ON STANDARDIZED TEST — ONE SOLUTION TO A 50 YEAR OLD PROBLEM

Analogy items (mean of six hard analogies compared with mean of six easy analogies)

		Adjusted Corrects (all guessing empirically removed)		**Total Guessing over 5 Options**	
Matched SAT Scores		**Blacks**	**Whites**	**Blacks**	**Whites**
200	**Easy**	26.78<	32.02	39.68>	29.82
	Hard	04.32=	04.32	59.92< **	59.72
210	**Easy**	32.79<	35.79	33.88>	31.03
	Hard	06.24>	03.84	57.07<	62.10
220	**Easy**	36.02<	40.04	30.46>	14.43
	Hard	06.52>	04.85	58.83<	62.65
230	**Easy**	38.74<	41.28	31.76>	22.53
	Hard	06.81>	5.01	67.56< **	65.44
240	**Easy**	42.92<	44.25	27.05>	23.18
	Hard	06.73> **	06.92	62.18<	64.07
250	**Easy**	44.44<	46.03	25.46>	21.20
	Hard	07.39>	06.70	63.82<	65.28
260	**Easy**	46.62<	46.77	22.35>	19.42
	Hard	08.35>	05.77	63.62<	67.06

[55] *Id.*

Note: Of 14 algebraically ordered predictions for "Adjusted Corrects," there was 1 violation (violations were indicated by **) and one equal sign. By a "sign" test, confirming twelve out of thirteen predictions provides support for the hypothesis at the p<.01 level of confidence. Of 14 algebraically ordered predictions for the columns labeled "Total Guessing over 5 options" there were two violations and twelve confirmed orderings. By a "sign" test, twelve confirmations out of fourteen predictions provides significant support for the hypothesis at the p<.01 level of confidence. For each SAT score, level equal numbers of White and Black examinees were selected (this was done to remove possible artifacts associated with different sample sizes prior to making ethnic comparisons). For SAT = 200 there were 737 Whites and 737 Blacks; for SAT = 210 there were 278 Whites and 278 Blacks; for SAT = 220 there were 626 Whites and 626 Blacks; for SAT = 230 there were 413 Whites and 413 Blacks; for SAT = 240 there were 530 Whites and 530 Blacks; for SAT = 250 there were 1,062 Whites and 1,062 Blacks; and for SAT = 260 there were 439 Whites and 439 Blacks. In all 8,085 low-scoring SAT students were analyzed. The reader should note that although there were a total of thirteen analogy items analyzed from SAT form QSA-09, the middle difficulty item (1-17) was removed to simplify the contrast between six "easy" analogies and six "hard" analogies.

Table 5 shows that overall the Whites are guessing MORE (at 63.76%) than matched-ability African-Americans (at 61.86%) on the hardest analogy items, and interestingly, that Whites are guessing LESS often (at 23.09%) than matched-ability African-Americans (at 30.09%) on the easiest analogy items. This result suggests several things. *The new guessing model shows that low scoring White and African-Americans are NOT guessing 100% of the time on the hardest items—rather, they are scoring at about the 64% and 62% levels.* The pattern of this result is consistent with my assertions in my 2003 article.[56] The new guessing model further shows that even on the easiest items, these low scoring students are still guessing, but at reduced levels. For the African-Americans, they are guessing about 30% of the time on the easy analogies while Whites are guessing 23% of the time. This is not a surprising result inasmuch as my 2003 paper has suggested that there is greater uncertainty of how African-Americans are probably interpreting the exact meaning of very

[56] *See generally* Freedle, *Correcting the SAT's*, *supra* note 6.

common vocabulary words, given that cultures can diverge quite widely on what such common words often refer to—for cultural groups that emphasize an extended family, common words such as "home" have a more ambiguous referent than it does for groups, such as Whites, that typically no longer have extended families.[57] An examination of the work of Diaz-Guerrero and Szalay shows that African-Americans and Whites differ strongly on such commonly used words as "justice," "progress," "society," and "class."[58] If your test materials involve minimal verbal context, such ambiguities begin to multiply. And this can explain why minorities have a differentially more difficult time correctly answering "easy" test items.

This new guessing model applies as well to guessing on the LSAT or any other standardized multiple-choice test—that is, this new model still applies regardless of whether students are penalized for guessing (as in the SAT) or not (as in the LSAT). Having a more accurate model of student guessing on standardized tests will help stop certain psychometricians from characterizing many minority examinees as just "lucky guessers." By solving a 50-year old problem, it will also help test developers fit their data more accurately no matter what the student demographics are.

IV. WHY THE LSAT IS PROBABLY ETHNICALLY BIASED

The LSAT is probably ethnically biased for the following reasons: (1) It leads to mean racial differences and (2) it probably contains significant individual item bias (i.e., Differential Item Functioning effects) which one can infer from earlier studies by Mary Enright and Isaac Bejar.[59]

The reader will recall one of the key inferences that resulted from our discussion of the Fagan and Holland studies. That is, if Fagan and Holland are correct in their assertion that there is no significant difference in verbal aptitude between African-American and White people, then any test that reports such

---

[57] *Id.* at 6–7.

[58] ROGELIO DIAZ-GUERRERO & LORAND B. SZALAY, UNDERSTANDING MEXICANS AND AMERICANS: CULTURAL PERSPECTIVES IN CONFLICT (1991).

[59] *See generally* MARY K. ENRIGHT & ISAAC I. BEJAR, EDUC. TESTING SERV., AN INVESTIGATION OF THE ROLE OF EDUCATIONAL BACKGROUND ON PERFORMANCE ON THE GRE ANALYTICAL MEASURE (1998), *available at* http://www.ets.org/Media/Research/pdf/RR-97-17-Enright.pdf.

mean differences must be culturally biased.[60] The LSAT does yield mean ethnic differences, therefore, by the above reasoning, it must be culturally biased.

There is other indirect evidence that the LSAT is biased when *individual test items* are examined. Most of the indirect evidence comes from a study by Enright and Bejar who examined DIF for the GRE.[61] What does a study of the GRE have to do with the LSAT? Well, other studies have shown that the GRE and the LSAT have a nearly identical factor structure,[62] which is to say that the kinds of items that the two tests use (e.g., the Analytical reasoning and Reading item types) yield mathematically similar factor results. Therefore, because the LSAT is most likely ethnically biased in both senses (mean ethnic differences as well as individual item level differences), use of LSAT test scores are contributing directly to decreasing minority enrollment in many law schools.[63] The partial solution, as I have suggested above, is at least to add the Fagan-Holland test of verbal aptitude as a required test for law school admission. Later, I will suggest an additional solution to the minority admission problem when we discuss the topic of how the invalid use of the LSAT has led, absurdly, to law school faculty ranking.

V. FACTORS THAT MAY AFFECT THE PREDICTIVE VALIDITY BETWEEN LSAT SCORES AND LAW SCHOOL GRADES

I would like to discuss how we can use the LSAT (along with the Fagan-Holland test) to improve our understanding of how *all* students—minority and majority alike—perform regarding the grades they receive in each of their three years in law school. That is, we can ask what specifically there is about the structure

---

[60] *See* Fagan & Holland, *supra* note 1, at 380 (finding that "differences in knowledge between Blacks and Whites for items tested on an intelligence test, the meanings of words, could be eliminated . . . when equal opportunity for exposure to the information to be tested had been experimentally assured").

[61] *See generally* ENRIGHT & BEJAR, *supra* note 59.

[62] *See* KENNETH M. WILSON & DONALD E. POWERS, LAW SCHOOL ADMISSION COUNCIL, FACTORS IN PERFORMANCE ON THE LAW SCHOOL ADMISSION TEST iii (1994), *available at* http://www.lsacnet.org/lsac/research-reports/SR-93-04.pdf ("[T]he study findings suggest a common underlying structure for logical reasoning, reading comprehension, and analytical reasoning item type regardless of the test (LSAT or GRE) in which they are used.").

[63] John Nussbaumer, Remarks at St. John's University School of Law Conference: The LSAT, *U.S. News & World Report*, and Minority Admissions (Sept. 7, 2005).

and content of especially the LSAT that accounts for its correlation with students' grades in the first, second, and third years of law school. This point therefore addresses the issue of the *predictive validity* of the LSAT along with other predictors such as undergraduate grades, a student's score on the Fagan-Holland test, and other factors to be described below. That is, ideally we want to understand, for example, why a high or low LSAT score predicts a high or low grade point average for each year in law school. We also want to know in broader terms whether coaching or mentoring can improve one's LSAT score, and whether mentoring can affect grades and graduation rates. Needless to say, changes in either LSAT scores and/or grade average will certainly affect the strength of the relationship between LSAT and grades earned in law school.

Some important earlier work done by Donald Powers examined the *changing* relationship between the LSAT and the grades earned for each of the three law school years.[64] Powers found that the LSAT correlates more strongly with grades for the first law school year and then systematically diminishes over each of the next two years.[65] In other studies, Powers also found that the grades earned by African-American law students *increase* systematically each additional year they are in law school.[66] White students' grades also improved over time, but not to the same significant extent that Black students' grades improved. Powers reported that "[i]n 18 of 21 law schools, Black students showed greater improvement than White students when third-year grades were compared with first-year grades. In 10 of 21 schools, the improvement of Black students was significantly greater, statistically, than that of White students."[67]

Within this set of results, let us add yet another finding. The first year of law school typically involves use of the Socratic method of instruction. The second and third years of law school

---

[64] *See* Donald E. Powers, *Long-Term Predictive and Construct Validity of Two Traditional Predictors of Law School Performance*, 74 J. EDUC. PSYCHOL. 568 (1982).

[65] *See id.* at 574.

[66] *See* DONALD E. POWERS, DIFFERENTIAL TRENDS IN LAW SCHOOL GRADES OF MINORITY AND NONMINORITY LAW STUDENTS (1982), *available at* www.ets.org/research/researcher/RR-82-21.html [hereinafter POWERS, DIFFERENTIAL TRENDS]; DONALD E. POWERS, LAW SCHOOL ADMISSION COUNCIL, PREDICTING LAW SCHOOL GRADES FOR MINORITY AND NONMINORITY STUDENTS: BEYOND FIRST YEAR AVERAGES (1984).

[67] *See* POWERS, DIFFERENTIAL TRENDS, *supra* note 66, at 1.

involve use of the "problem method" of instruction. What is interesting about the "problem method" is that students come to class knowing in advance what problem will be addressed. As a consequence, there is typically not the same time pressure to come up with a snap answer in the second and third years of study as there is in the first year of study.

If we put this first group of facts together, it suggests that the predictive validity of the LSAT diminishes over successive years because the time students have to prepare for classroom discussion is distinctly different for especially the first year of study in contrast with the subsequent years. Of course, this differential effect of time pressure and its possible effect on the predictive validity between LSAT scores and grades are reminiscent of William Henderson's work which suggests that the LSAT involves a speeded component.[68] Henderson reasoned that grades which specifically depend upon take-home exams or essay papers—student work which does *not* involve a strong speeded component—should be less strongly correlated with LSAT scores than in-class exams that do involve a speeded component.[69] I believe this is precisely what he found.

Henderson's work is significant because it shows that careful analysis of the critical components that go into the LSAT test situation and the components that are involved in the criterion setting (the types of school work that get graded) together help us better understand the magnitude of the relationship between criterion and test. Along this line of reasoning, I would further like to suggest that one take the total LSAT score *and break it into at least two subscores: the analytical reasoning subscore and the reading subscore.* It is possible that there are some law course grades that are differentially sensitive to either the reasoning subscore or the reading subscore. Carrying out such a study with and without special mentoring activities regarding just the reading or reasoning components should further improve our understanding of what predictive validity is all about. I would also suggest that adding a race-free measure of verbal

---

[68] *See* Henderson, *supra* note 2, at 979 (presenting empirical evidence that test-taking speed is "a variable that affects student performance on both the LSAT and actual law school exams").

[69] *See id.* at 995 ("[T]he strength of correlation between the LSAT and law school grades may vary in proportion to the number of grades that are determined by speeded, in-class exams.").

aptitude such as the Fagan-Holland test into this mix would further increase our understanding of the critical components of law school grades.

You will recall I mentioned above that Powers found that African-American law students show an accelerated increase in grade point average over the three years in law school. Are there certain law school courses that account for most of this increase? Is there a differential effect of LSAT *subscores* for these special law school courses that would help explain this dramatic increase? There is obviously need of much additional work.

Not all LSAT studies are in agreement with the Powers and Wilson works. For example, Linda Wightman conducted a study of the relationship between LSAT scores and grades for each of the three law school years and failed to find a different trend in the strength of the relationship between LSAT scores and grades over the three years of law training.[70] Wightman's LSAT test was almost certainly different in its subsections (especially in the type of items designated as Analogical Reasoning) from the LSAT studied earlier by Powers. It may be these structural and content differences in the LSAT itself that account for the different findings. Again, more research is needed to clarify these conflicting findings.

I mentioned above that Wilson's[71] work supports the general findings of Powers[72] regarding the finding that minority students continue to improve as their grades accumulate over their years of study. However, Wilson's work was restricted to undergraduate performance. What I want to highlight here is that Wilson refers to this phenomenon as evidence of "late blooming" among minority students in college.[73] Bowen, Kurzweil, and Tobin refer to a similar phenomenon in their book,

---

[70] *See* LINDA F. WIGHTMAN, LAW SCHOOL ADMISSION COUNCIL, BEYOND FYA: ANALYSIS OF THE UTILITY OF LSAT SCORES AND UGPA FOR PREDICTING ACADEMIC SUCCESS IN LAW SCHOOL 37–38 (2000).

[71] *See* KENNETH M. WILSON, PREDICTING THE LONG-TERM PERFORMANCE IN COLLEGE OF MINORITY AND NONMINORITY STUDENTS: A COMPARATIVE ANALYSIS IN TWO COLLEGIATE SETTINGS (1980) [hereinafter WILSON, PREDICTING]; *see also* Kenneth M. Wilson, *Analyzing the Long-Term Performance of Minority and Nonminority Students: A Tale of Two Studies*, 15 RES. HIGHER EDUC. 351, 368 (1981) [hereinafter Wilson, *Two Studies*].

[72] *See* POWERS, DIFFERENTIAL TRENDS, *supra* note 66.

[73] *See* WILSON, PREDICTING, *supra* note 71; *see also* Wilson, *Two Studies*, *supra* note 71, at 368 (discussing comparatively high across-class minority gains in GPA).

Equity and Excellence in American Higher Education.[74] That is to say, minority students and other disadvantaged students (e.g., students from the lowest economic level) who are given an opportunity to attend elite colleges often rise to the occasion, graduating at a substantial rate.[75] For example, the percentage students from the lowest income quartile graduated at 84.4% while those at the highest income level graduated at 87.6%.[76] After correcting for other variables such as differences in SAT scores, race and so forth, the adjusted graduation rates were 80.9% for the lowest income quartile and 85.6% for the highest income quartile. So, in spite of coming from disadvantaged backgrounds, these students in the lowest income quartile managed to graduate at an encouragingly high rate relative to students at the highest income levels who started their education with all the advantages of great wealth and attendance at excellent preparatory schools.[77] Bowen et al. also briefly discuss how mentoring programs affect the success of minority students enrolled in higher education. They indicate that "only 18 months after being matched with mentors, student participants were less likely to engage in self-destructive behavior or skip classes, had higher average grades, and felt more confident of their academic abilities."[78] Clearly, mentoring can alter a student's grade point average which, in turn, is likely to affect the magnitude of the predictive validity relationship between grades and, presumably, in the case of law school, the LSAT scores.

Finally, we need to consider whether Adam Fisher's method of improving LSAT scores—as discussed in an article by Hope Reeves in a recent *New York Times* piece—uncovers hidden levels of ability or motivation within students.[79] What are these confusions that Fisher manages to erase? More broadly, how might mentoring (ala Fisher) or other coaching techniques differentially affect the level of law school grades achieved and/or the level of LSAT performance achieved? Without mentoring,

---

[74] WILLIAM G. BOWEN, MARTIN A. KURZWEIL & EUGENE M. TOBIN, EQUITY AND EXCELLENCE IN AMERICAN HIGHER EDUCATION (2005).

[75] *See id.* at 119–20.

[76] *See id.* at 120 figs.5.11a & 5.11b.

[77] *See id.*

[78] *Id.* at 241–42 (footnote omitted).

[79] *See generally* Hope Reeves, *Tutors Hold Key to Higher Test Scores, for a High Fee*, N.Y. TIMES, June 1, 2005, at B10 (detailing Adam Fisher's tutoring method for the LSAT).

"stereotype threat" effects during LSAT testing or during classroom testing might remain at a high level.[80] That is, minority students who feel threatened may fail to show up for many classes, leading to lower grades. But after mentoring or coaching, stereotype threat may well be mitigated with greater classroom attendance likely, therefore yielding higher classroom grades (from the classroom testing component). Again, as Henderson has suggested, we need to specify how many components are shared by the law professor's grading method, and the elements of the LSAT itself, if we are to understand why the predictive validity correlation reaches the level (typically .40) that it does.[81] Once we know what the critical components are, perhaps the student mentoring sessions can focus more intensively on precisely these common factors in order to at least boost law school grades.

VI. WHY THE LSAT SCORES OF LAW STUDENTS AS USED IN THE ANNUAL RANKINGS OF LAW SCHOOLS BY *U.S. NEWS & WORLD REPORT* ARE AN INVALID TOOL FOR EVALUATING THE QUALITY OF LAW SCHOOL FACULTY INSTRUCTION: SUGGESTIONS FOR IMPROVING AND EXPANDING THE RANKING SYSTEM TO MAKE IT REGIONALLY MORE USEFUL AND TO LEAD TO A RANKING OF FACULTY QUALITY DISTINCT FROM THAT OF STUDENT-RELATED ISSUES

Some background information is necessary before we begin. I found it quite surprising to learn that *U.S. News & World Report* is ranking the quality of law school education by using primarily the incoming students' LSAT scores (i.e., the correlation between the existing law school rankings and the LSAT scores is extremely high, about .80 based on some re-analyses I have conducted on my own). So, the incoming students' LSAT scores are dominating the overall law school rankings. Such a ranking system is not only absurd, but it also has a very corrosive and negative effect on minority law school enrollments across the country.[82]

Logically, how can the quality of education offered by the

---

[80] *See* Steele & Aronson, *supra* note 18, at 401.

[81] *See* Henderson, *supra* note 2, at 1008–09.

[82] *See* John Nussbaumer, *Misuse of the Law School Admissions Test, Racial Discrimination, and the De Facto Quota System for Restricting African-American Access to the Legal Profession*, 80 ST. JOHN'S L. REV. 167, 167–68, 170 (2006).

faculty of an educational institution be primarily dependent upon the test scores of an incoming student body? It simply is not a defensible position to maintain. Another illogical aspect of these rankings has to do with the following facts. *U.S. News & World Report* also ranks the law schools for their expertise in each of nine specialties (e.g., International law, Tax law, Environmental law, etc.). *Another fact is that Yale Law School is ranked Number 1 in the nation, yet, when one scans which law schools rank among the top three positions for each of the nine specialties, Yale is nowhere to be found*! How is this possible? How could the best law school in the country not excel in any of the nine specialties? The reason this is possible, I maintain, is that the overall national ranking is strongly contaminated with the LSAT scores of incoming students, while the ranking for the nine specialties is presumably based on actual faculty merit within each field. Apart from the fact that this type of inconsistency is an embarrassment, the real question is: What other type of ranking or rankings might make greater sense? Below I suggest a solution based on the idea that a clear separation of rankings based on faculty merit apart from rankings, which are germane to student issues, should be introduced.

I just mentioned that *U.S. News & World Report* presents the rankings of the top law schools in each of nine specialties: (1) clinical training, (2) dispute resolution, (3) environmental law, (4) healthcare law, (5) intellectual property law, (6) international law, (7) legal writing, (8) tax law, and (9) trial advocacy. I first suggest constructing a new ranking system for law school quality based upon a weighing of a school's standing with respect to all nine of these specialties—to make this idea crystal clear, a numerical illustration of the procedure is provided below.

Before I illustrate the procedure, two additional ideas need to be introduced. Professor Vernellia Randall independently suggested that a regional ranking of law schools would lead to more useful information for prospective law students.[83] As a consequence of Randall's suggestion I modified my idea of a national re-ranking of law schools (obtained by merging each school's relative standing across all nine specialties) to reflect a

[83] Vernellia Randall, Remarks at St. John's University School of Law Conference: The LSAT, *U.S. News & World Report*, and Minority Admissions (Sept. 7, 2005).

law school's regional standing, again with respect to the merging of each school's relative standing across all nine specialties within a given region. Professor Leonard Baynes, in subsequent correspondence, further suggested that Civil Rights be added as a tenth specialty. Adding Civil Rights as a specialty might serve to increase admissions officers' interest in enrolling more minority students who clearly have a special interest in this specialty.

I find Randall's suggestion to be important for several reasons. By focusing on one specialty at a time for a limited region of the country, the people conducting the ranking have a cognitively more manageable task than the national ranking approach currently used by *U.S. News & World Report*.[84] Such a cognitively more manageable task should lead to greater reliability in the rankings.

I now illustrate this new ranking method with a simple numerical example. We wish to find a school's final ranking within a region (here, just the Northeast) by summing "points" earned by that school across all the specialties. To keep things simple, suppose there were only four law schools in the entire Northeast region: Harvard, Yale, Columbia, and NYU. Further, suppose that there were only two specialties of law called A and B. Let's say that for *specialty A* the ranking for the four schools is: Harvard (rank 1 = 4 points),[85] NYU (rank 2 = 3 points), Yale (rank 3 = 2 points), and Columbia (rank 4 = 1 point). For the second *specialty B*, the rankings of these same four Northeast schools is: Columbia (rank 1 = 4 points), Harvard (rank 2 = 3 points), Yale (rank 3 = 2 points), and NYU (rank 4 = 1 point). Now to find the overall ranking across the specialties in the Northeast region we re-rank the four schools according to the total number of points they have earned across all the specialties. This leaves us with the following result: Harvard (rank 1 with 7 points), Columbia (rank 2 with 5 points), and NYU and Yale are tied for rank 3 (each getting 4 points). Such a regional ranking, according to the relative standing of each law school with respect to the specialties in the profession, is a clear reflection of overall

---

[84] One would be ranking, say, at most, 50 law schools in a given region rather than 180 schools across the nation.

[85] The reader should note that a maximum of four points exists here only because we have restricted the number of schools to 4. If there had been 50 schools in this region, the maximum number of points would have been 50.

law school proficiency within a given region of the country. As such, these specialties have nothing at all to do with how high (or low) their incoming student body's LSAT scores are. Because the specialties are logically unrelated to LSAT scores, a student who examines these regional rankings should ideally be attracted to those schools that offer the best training in the specialties of greatest interest to the student.[86] A *separate ranking* should be provided which reflects the special concerns of student applicants: such as the number of scholarships offered, externship options, ethnic and gender diversity, cost, alumni support network, the mean LSAT scores of other students with whom they will study, the quality of library holdings, etc. The main point here is that I recommend clearly separating a law school faculty's status ranking from any ranking associated with student LSAT scores.

Robert Morse, the director of research for *U.S. News & World Report* provided comments that seem to suggest that there will be no substantial changes in how rankings are formulated. He stated that "[L]aw school rankings are here to stay."[87] Furthermore, the consumers of his magazine find these rankings useful. But if the *U.S. News & World Report* magazine is not interested in changing its current ranking methodology to achieve a more rational basis, perhaps *other* national magazines would be interested in developing and publishing such a new approach. Or, perhaps some enterprising student internet group might supply the newer, more rational rankings for free! With readership down for published materials and internet readership up, this might well be the most effective future antidote for replacing the current questionable and inconsistent rankings with, what I would judge to be, fairer ones.

Mr. Morse, in his further comments, is technically correct in reminding his audience that his magazine "is not responsible for the use of the LSAT in the admissions process" and that "law schools are the ones that determine who is admitted."[88] (In other words, if you purposely set a fire, you can always blame the fire

---

[86] Incidentally, it should be obvious that a *national* re-ranking of law school faculties across the several specialties is also entirely feasible using the above method of assigning points within each specialty.

[87] Robert Morse, Remarks at St. John's University School of Law Conference: The LSAT, *U.S. News & World Report*, and Minority Admissions (Sept. 7, 2005).

[88] *Id.*

department for not putting out the blaze soon enough.) Nevertheless, the admissions officers are blameworthy for their slavish response to these annual law school rankings. I find this fact in itself astounding. How can powerful *law* schools allow themselves to be so easily victimized by what a magazine publishes? And why are they so helpless in offering up solutions? Is the law so weak, are precedents so wanting, that they cannot summon or fashion a rule by which to turn this dilemma into victory? As I have said in my opening pages, if the Fagan-Holland test were to replace the LSAT as an assessment of verbal aptitude, some of these rancorous issues would soon disappear. But I would be surprised if such logic would win many converts.

Professor Randall, in the speech she was invited to give at this conference, has indicated that law schools might be vulnerable to litigation if they continue to restrict the admission of minority students into their schools based, at least in part, on the minorities' LSAT scores. To which I would add, is this not strangely ironic that our very own *law schools* may need the coercion of the law to make them honor the recent Supreme Court's ruling to re-affirm Affirmative Action?[89]

---

[89] The *U.S. News & World Report* has just published its 2006 edition of graduate school rankings. U.S. NEWS & WORLD REPORT, AMERICA'S BEST GRADUATE SCHOOLS 2006 (2005). I have analyzed their new rankings of the top 102 law schools and examined the correlation with LSAT (75th percentiles) scores. The new data yield an extremely large correlation (r = -.88 which is highly significant) indicating that school rank is increasingly tied to LSAT scores. But do the rankings of other graduate school programs also produce such strong relationships with student scores on standardized tests? Here are the results: For Medical Schools (Primary Care), the correlation of rank with MCAT test scores was only -.19. (This is not even statistically significant!). For Graduate Engineering schools, the correlation of rank with GRE-math was only -.43. While this is statistically significant, it represents a weak relationship. For Graduate Education, rank again correlated -.43 with GRE-math, but separately correlated a somewhat higher -.55 with GRE-verbal. For Medical-Research, rank correlated -.75 with student MCAT scores. And finally, for Graduate Business School, rank correlated -.85 with student GMAT scores. This last finding between rank and student test scores is the only one that clearly rivals LSAT and school rank in terms of the strength of the relationship. In order to double-check that the LSAT correlation was not somehow sensitive to the much larger sample of 102 schools, I re-computed the rank/test-score correlation for just the top 51 law schools to make it more comparable in size to the other graduate school results. The results still produced a very strong correlation of -.89 between school rank and student LSAT scores. So, sample size does not explain the large LSAT correlation with school rank. Such remarkable variability in the strength of the relationship between school rank and student standardized test scores across these several graduate school disciplines needs to be explained. Why does law school ranking yield an almost perfect relationship between rank and student scores whereas several of

In summary, I recommend clearly separating faculty ratings from student ratings. I furthermore recommend ranking by region (as suggested by Randall). I further recommend ranking law faculty not only within each region but especially for their cumulative excellence across each of several specialties, including Civil Rights (as suggested by Baynes). I also recommend, as stated earlier, a separate ranking of schools based on a variety of *student services* that are provided (such as scholarships, alumni support network, cost, externship options, ethnic and gender diversity, LSAT scores of other students at the school, etc.).[90]

CONCLUSION

A final word: Tests are omnipresent in our complex society. For all their presumed benefits (e.g., greater efficiency in processing large numbers of people, etc.), a mindless application of test results can also be dangerous to a progressive, democratic society. For example, tests can be powerfully negative determinants of how our schools function (e.g., "Teach to the test"). Tests can influence who gets hired (many businesses now require disclosure of an applicant's SAT scores). Test results influence racial theories of genetic superiority and inferiority.[91] Tests can be erroneous in the sense that they can fall far short of their intended beneficial goals.[92] Tests can distort the true ability of large groups of disadvantaged students.[93] Tests can be

---

the other graduate school programs yield a much weaker relationship? Is this solely the responsibility of the *U.S. News* research staff or are the law school admission officers, by exercising special standards in selecting a new student body, purposely contributing to this extraordinarily close tie between test scores and school rank? Not all graduate school programs provide test scores nor do they always provide grade-point-averages as the new 2006 survey makes evident. So why do law schools? It seems to me that minorities would have much to gain if LSAT scores as well as grades were to be withheld from the *U.S. News & World Report* magazine team since their law school rankings would then have to be based on something other than a heavy reliance on especially LSAT scores. Yet, when some graduate programs do provide test score data, it does not invariably strongly influence school ranking. The answer to this last puzzle needs to be unraveled in a careful step-by-step analysis.

[90] *See* LAW SCHOOL ADMISSION COUNCIL, LAW SCHOOL DEANS SPEAK OUT ABOUT RANKINGS (2005), http://www.lsac.org/pdfs/2005-2006/RANKING2005-newer.pdf.

[91] *See* ARTHUR R. JENSEN, BIAS IN MENTAL TESTING 2–3 (1980).

[92] NICHOLAS LEMANN, THE BIG TEST: THE SECRET HISTORY OF THE AMERICAN MERITOCRACY (1999).

[93] *See* Freedle, *Correcting the SAT's*, *supra* note 6, at 1–7 (arguing that the SAT is both culturally and statistically biased, as well as a poor determinant of the

used for invalid purposes and can be seriously corrosive (e.g., evaluating law faculty based on student LSAT scores). But the good news is that at least some tests can be uplifting and strongly beneficial, such as the Fagan-Holland test of racial equality. Yes, tests are here to stay and tests will continue to be consumed by many unguarded consumers. To which I would simply add, "Caveat emptor."

---

academic abilities of minority students).