Promoting Equitable Law School Admissions through Legal Challenges to the LSAT

Al Alston
Promoting Equitable Law School Admissions through Legal Challenges to the LSAT
Alson Alston
Dickinson School of Law
Juris Doctor Candidate 2015

We can countenance a law suit to be fought within the next six months in which the lack of scientific evidence will win the jury or convince the defendant(s) to settle. The plaintiff will challenge a law school’s denial of her application for admission. That denial will be based, in part, on the applicant’s LSAT score(s).

“How large an effect time limits have on the LSAT is not currently known because we cannot identify rapid-guessing behavior.”
Law School Admission Council Researchers
Table of Contents
1  Introduction ................................................................................................................................. 1
2  Speededness, Validity and Reliability ......................................................................................... 2
   2.1  Defining speededness ............................................................................................................. 2
   2.2  Identifying speededness – LSAT’s incomplete, inconsistent and impossible data collection and classification problems .................................................................................................................. 5
   2.3  Models of speededness do not rescue LSAT’s data collection and classification failures........ 8
      2.3.1  Gulliksen: Theory still accepted, implementation fails under assumptions and data classification problems -- Assumes no omissions, examines end-of-test only, ignores rapid-guessing 8
      2.3.2  The Educational Testing Service’s (ETS’s) floor/ceiling model and LSAC’s per cent model 11
      2.3.3  All other methods of identifying speededness are fatally flawed for pencil-and-paper exams 14
      2.3.4  Computer-Based testing .................................................................................................. 17
   2.4  The effects of speededness ..................................................................................................... 19
      2.4.1  A test is speeded if and only individuals find it speeded: Reasserting the real definition of speededness .......................................................................................................................... 19
      2.4.2  The impacts of speededness on reliability and validity .................................................. 25
   2.5  Conclusion ............................................................................................................................. 32
3  Legal Action against LSAC and Other Parties ........................................................................... 34
   3.1  Injury caused by the LSAT ..................................................................................................... 34
   3.2  Liability for injuries caused by the LSAT ............................................................................. 35
   3.3  Heart of the lawsuit ............................................................................................................... 36
   3.4  Possible Counts I, II and III of lawsuit .................................................................................. 37
   3.5  Daubert hearing – Admission of LSAC’s scientific evidence is unlikely ................................. 39
4  Conclusion ................................................................................................................................... 41
   4.1  Caveat .................................................................................................................................... 42
5  Appendix A: Harold Gulliksen’s Speededness Model ................................................................. 44
6  Appendix B: Deriving Test Speededness from Data Measuring Individual Speededness............ 49

Tables
TABLE 1. CYCLE OF LSAC TECHNICAL REPORT TOPICS ............................................................... 27
TABLE 2. LSAT PREDICTIVE VALIDITY STUDIES, FROM 1995 – 2008 ........................................ 31
1 Introduction

A preponderance of the available evidence demonstrates that the Law School Admission Council (LSAC) cannot prove that the LSAT is a valid and reliable measure of ability because LSAC’s time allotment for the exam, its standards for classification of speededness, and its very measurement of speededness are arbitrary. In fact, the measurement of speededness is impossible under the current testing conditions. LSAC’s arbitrary standards for classification of speededness, therefore, which admittedly tell us little about which thresholds allow meaningful interpretation of test scores, are applied against mere conjecture as to the speededness actually present. Interpretation of test scores under these circumstances is speculative.

And, yet, by even its arbitrary speededness standards and measurements, some LSAC researchers have demonstrated that the test is speeded for African-Americans and Puerto Ricans, while not speeded for others. More careful measurements of speededness indicate that the LSAT might be speeded for 50 per cent of all test-takers. The errors introduced into the validity and reliability of test scores, when the levels of speededness observed in the LSAT occur, lead to numerous destructive inferences about the LSAT. Therefore, even accepting LSAC’s speededness standards and measurements, there is evidence that the test is invalid and unreliable for African-Americans, Puerto Ricans and, possibly, half of all test-takers.

The key players within the law school industry either know or should know of the serious issues facing the LSAT. Those players include LSAC, the creators of the LSAT; law schools, which continue to require applicants to take the test and make admission decisions, in

---

1 This section is based upon the facts and analysis presented later. Therefore, it is footnoted only where absolutely necessary, while the rest of the discussion liberally cites to all sources used.

2 Speededness will be discussed, in detail, below. For now, please consider it to be the inability of a test-taker to consider fully all questions on a test, due to an inadequate time limit.
part, on LSAT scores; the American Bar Association (ABA), which accredits virtually every law school in the United States, pursuant to a contract with the U.S. Department of Education (DOE). *U.S. News* is also a part of the industry, due to its yearly rankings, which include LSAT score ranges for each law school. These stakeholders all share liability in a scheme that has, for decades, judged test-takers unworthy of entry into one or more law schools through guesswork, self-serving metrics and research, and a false cloak of credibility.

Below, we explain the concept of speededness. We show how LSAC has defined and measured it, why such definitions and measurements are unjustifiable, how the errors LSAC has traditionally introduced impact inferences to be drawn from test scores. We demonstrate that LSAC cannot prove the validity and reliability of the LSAT. We then briefly discuss the injury caused by the LSAT and the areas of law where a test-taker could prove liability.

## 2 Speededness, Validity and Reliability

### 2.1 Defining speededness

Educational testing research identifies two broad categories of tests: speed and power.³ In a speed test, the score is determined by the number of questions attempted, while the measure in a power test is the number answered accurately.⁴ Researcher Harold Gulliksen defined the pure speed test as one that was so easy, every test-taker would answer every question correctly, but it was so long that no test-taker could finish it.⁵ A pure power test is a test in which all items are attempted after full consideration, so performance is measured by

---

⁴ Id.
the number of correct answers. Thus, a pure speed test is one where scoring is based on rapidity of response, while a pure power test is based on ability.

Understanding the construction and scoring of a pure speed or pure power test would appear to be a simple matter. However, standardized testing that ostensibly seeks to measure ability has traditionally been implemented at testing centers with time limits present, largely, for administrative and practical purposes. The introduction of time limits into a power test raises significant issues regarding the method and interpretation of scoring. To conceptualize the introduction of time into otherwise power tests, standardized testing researchers use the term “speeded” or “speededness.”

LSAC researchers Deborah L. Schnipke and David J. Scrams define speededness as “the extent to which time limits affect test takers' performance.” The very fact that time limits are imposed on an alleged test of ability forces some, if not all students to work faster than they might in an untimed test so that they can answer as many questions as possible. The dilemma for test administrators is that if they set a time limit too short, then they undermine the validity of what is constructed and reported to be a power test, because few of the test-takers will have an opportunity to complete all of the questions. If they set the time limit too

---

6 Lu & Sireci, supra note 5, at 29.
7 Id.
8 Lu & Sireci, supra note 5, at 30.
9 Rindler, supra note 3, at 261.
10 Id.
12 Lu & Sireci, supra note 5, at 29.
13 Id.
long, then early finishers might disrupt the others or the test could become difficult to administer.\textsuperscript{14}

Clearly, in a pure power test, as LSAC asserts that the LSAT is,\textsuperscript{15} any speededness would invalidate the basic assumptions of the test – that the test-taker fully considered the question before answering.\textsuperscript{16} Precisely because alleged power tests are not pure, the issue for test administrators becomes how much speededness is acceptable and how much invalidates the test or makes the test unreliable.\textsuperscript{17}

The great challenge in 1950, when Guilford was a pioneer in the science of educational testing, was that most standardized tests were a combination of both speed and power, so obtaining a valid scoring method was unclear:\textsuperscript{18}

Relatively little has been written on [distinguishing between speed and power tests], despite the fact that the problems of item analysis, test length, item difficulty distribution, determination of reliability, and error of measurement are all quite different for the two types of tests. At present, most tests are a composite in unknown

\textsuperscript{14}Id.
\textsuperscript{15}LSAC, \textit{Cautionary Policies Concerning the LSAT}, LSAC, \url{http://www.lsac.org/docs/default-source/publications-(lsac-resources)/cautionarypolicies.pdf} (last visited Dec. 10, 2013). LSAC states:

\begin{quote}
LSAT scores provide a standard measure of an applicant’s proficiency in the well-defined set of skills included in the test. \textit{Id.} at 1.
\end{quote}

\textit{See also} Schnipke & Scrams, \textit{supra} note 11, at 2. Their report is an LSAC report, so it must be considered an authoritative statement of the fact that LSAC, at least through its public face, designs the LSAT to be a power test that suffers from speededness for administrative reasons only:

\begin{quote}
Most aptitude and assessment tests are designed essentially as power tests, but a speed component is introduced when these tests are administered with a time limit-a practice often adopted for group administration. The Law School Admission Test (LSAT) is such a test. \textit{Id.}
\end{quote}

Schnipke & Scrams further add that speededness is present on the LSAT as a byproduct, so LSAC must learn how to measure it accurately:

\begin{quote}
The primary function of the LSAT is not to measure rate of work (speed), thus speededness is considered an ancillary variable. To know how great an effect this ancillary variable has on test scores, an accurate measure of speededness is needed. \textit{Id.}
\end{quote}

They note that the very purpose of their study is to obtain accurate measurements of speededness should the LSAT move to computer-based administration, because LSAC’s current means of measuring speededness for pencil-and-paper test administrations is simply an “impossible” task. \textit{Id.} at 16.

\textsuperscript{16}Lu & Sireci, \textit{supra} note 5, at 29.
\textsuperscript{17}Id.
\textsuperscript{18}Id.
portions of speed and power, which makes the development of appropriate theorems in test theory more difficult than for the pure type tests. (p. 230)\textsuperscript{19}

Unfortunately, little has changed in sixty years, because researchers today still have precisely the same difficulties:

We believe this is still the case today. Most educational achievement tests do not include speed of response in their definition of the construct measured and so it can be assumed they are considered to approximate a pure power test. Unfortunately, many test publishers do not report analyses of test speededness in their technical manuals or the results of empirical studies that were used to ensure time limits flexible enough to ensure speededness does not affect students’ performance negatively.\textsuperscript{20}

Of course, to determine the degree of speededness and whether that degree invalidates the test, one must first identify criteria for measuring speededness and identify standard criteria or boundaries for ensuring validity and reliability – if they exist.

\subsection*{2.2 Identifying speededness – LSAT’s incomplete, inconsistent and impossible data collection and classification problems}

For the LSAT, given only via pencil and paper,\textsuperscript{21} and similar administrations, the traditional identification technique has been to examine the number of questions never reached by the test-taker.\textsuperscript{22} LSAC researchers Schnipke & Scrams offer that:

With regard to the Law School Admission Test (LSAT), speededness is currently measured by calculating the proportion of test takers who do not reach each item on the test. These proportions typically increase slightly toward the end of the test ....\textsuperscript{23}

\begin{flushleft}
\begin{footnotesize}
\textsuperscript{19} Id., quoting Gulliksen, supra note 5, at 230.
\textsuperscript{20}Lu & Sireci, supra note 5, at 29.
\textsuperscript{21}LSAC, The LSAT: Day of the Test, LSAC, \texttt{http://www.lsac.org/jd/lsat/day-of-test} (last visited Nov. 30, 2013). The instructions provided by LSAC include: \textbf{Pencils.} Bring to the test center three or four sharpened No. 2 or HB wooden (for Canadian centers) pencils—with good erasers. \textbf{Mechanical pencils are prohibited.} All answers must be recorded on your LSAT answer sheet in pencil. Pencils must be used for the writing sample as well. \textbf{Pencils or sharpeners will not be supplied at the test center.} A pencil and a highlighter may be used to underline passages in the test book. Ink or ballpoint pens are not permitted. (emphasis original)
\textsuperscript{22}Rindler, supra note 3 262, 265-66.
\textsuperscript{23}Schnipke & Scrams, supra note 11, at 2.
\end{footnotesize}
\end{flushleft}
In fact, there are other, more accurate means of determining speededness than looking only at the end of a test. Test-takers may skip questions, without fully considering them, so that they can attempt later questions. These omitted questions, similar to those at the end of an exam, are never actually reached, so should appear in any speededness calculation. As well, some questions that are answered are not always fully considered by the test-taker, as a test of ability would otherwise suggest. On the LSAT, for instance, there is no penalty for incorrect answers, so a common test strategy is to engage in rapid guessing as time begins to expire. Such behavior may be more or less helpful, depending on the algorithm used on a particular test. Of course, some test-takers are not aware of this strategy, or otherwise do not employ it, leaving items blank.

These more sophisticated indicia of speededness are difficult or impossible to detect using either a single test or a parallel test administered with pencil and paper because no timing data is supplied for each question. Some researchers attempt to differentiate between

---

24 Id. at 265-66.
25 Id.
26 Id.
27 Id. at 3.
28 Although fairly deducible independently, this is an inference to be drawn from I. I. Bejar, Test Speededness under Number-Right Scoring: An Analysis of the Test of English as a Foreign Language (Report No. ETS-RR-85-11), ETS Res. Rep., (ETS. Newton, PA), 1985, at 1, 3. The suggestion is:

Although it is not pretended that the theory describes different strategies students may use in going through the test, it is one possible explanation for the lack of monotonicity that are observed sometimes with difficult items. Id.
29 Rindler, supra note 3, at 266.
30 Id. at 265-69.
31 See Schnipke & Scrams, supra note 11, at 16. They use the words “not possible” in their description of pencil-and-paper administrations:

With response times available, methodology such as that developed in the present study can be used to detect rapid-guessing behavior. This is not possible on paper-and-pencil administered tests. How large an effect time limits have on the LSAT is not currently known because we cannot identify rapid-guessing behavior. Id.
rapid-guessing behavior at the end of an exam, for instance, and incorrect answers resulting from lack of knowledge, by comparing the relative performance at the end of the exam to that of earlier parts of the exam.  If there is a significant improvement or decrease, then the researcher infers rapid guessing was the reason. For this to be true, however:

1. The exam would have to be uniformly difficult, so that apples-to-apples comparisons are being made;
2. There could not have been rapid-guessing behavior elsewhere in the test;
3. The type of rapid-guessing strategy employed by a given test-taker must have been likely to result in a different performance than non-guessing behavior for that test-taker; etc.

These conditions are extremely unlikely to be present on a consistent basis, if ever. It would similarly be difficult or impossible to differentiate between: questions not answered due to time constraints and not answered after full consideration, questions not fully considered due to time constraints versus those not fully considered due to nervousness, distraction or other matters not time-related; etc.

It should be clear, at this point, that LSAC has never identified speededness with any degree of accuracy when looking only at end-of-exam unanswered questions. Neither can

See Lu & Sireci, supra note 5, at 34 for limited discussion on this point. Rindler’s extensive discussion of mischaracterization and poor assumptions with respect to “unreached” questions, omissions, and rapid-guessing behavior readily leads to this inference, as well. See Rindler, supra note 3, at 265-69.

See Bejar, supra note 28, at 5; Rindler, supra note 3, at 268.

Id.

See the assumptions throughout Bejar, supra note 28, at 1-10; Rindler, supra note 3, at 266-69.

Id. See also Rindler, supra note 3, at 268-69.

Rindler, supra note 3, at 268-69. See also Schnipke & Scrams, supra note 11, at 16 (“How large an effect time limits have on the LSAT is not currently known because we cannot identify rapid-guessing behavior.”). Any class action suit against LSAC could identify the degree of culpability and pervasiveness of injury, in part, by examining
LSAC identify speededness with accuracy when considering the other indicia of speededness, because it cannot reliably differentiate between time-relevant reasons and ability-centered reasons.\textsuperscript{37} Thus, LSAC’s data collection problem is that it cannot and has not obtained response-time data for each question it poses to its test-takers, given its pencil-and-paper test administrations. LSAC’s data classification problem is that it cannot differentiate between the types of response data it does collect, due, primarily, to the absence of response-time data.

2.3 Models of speededness do not rescue LSAT’s data collection and classification failures

In an effort to determine the impact of LSAT’s data collection and classification problems and the imprecision arising from speededness, we examine the models used to characterize and measure speededness for the collective set of test-takers. Since this is legal research paper and not a statistics report, we will limit the presentation of formulas and place much of this limited set in appendices, though a significant amount of mathematics is essential to demonstrate key points. We will note why researchers have concluded that most of the models themselves tend to be poor characterizations of LSAT speededness, and fail utterly to measure speededness in light of the LSAT’s data collection and classification problems.

2.3.1 Gulliksen: Theory still accepted, implementation fails under assumptions and data classification problems -- Assumes no omissions, examines end-of-test only, ignores rapid-guessing

Harold Gulliksen provided what is considered to be the classic theory of speededness and both his assumptions and techniques are still in use today, for better or worse.\textsuperscript{38}

---

\textsuperscript{37} This is the ultimate conclusion of Rindler about any standardized test administered with pencil and paper.

\textsuperscript{38} Rindler, \textit{supra} note 3, at 262.
Gulliksen explicitly assumed that test-takers did not skip questions (defined as omissions), so that the only unanswered questions were at the end of the exam.\(^{39}\) Adopting the notation of Rindler (which actually corrects for some of Gulliksen’s assumptions), Gulliksen’s model can be expressed using:

\[
\begin{align*}
R &= \text{items correctly answered} \\
W &= \text{items incorrectly answered or omitted after consideration} \\
U &= \text{items unattempted (not considered, not reached within the time limit)} \\
N &= \text{the total number of items} \\
E &= \text{items NOT marked correctly (unattempted, incorrect, omitted after consideration)}
\end{align*}
\]

therefore,

\[
E = W + U \\
R = N - E^{40}
\]

Taking the standard deviation, \(S\), of each side of the equation \(E = W + U\), yields the following relationship between items NOT marked correctly and items unattempted:\(^{41}\)

\[
S_{(W+U)}^2 = S_W^2 + S_U^2 + 2r_{WU}SWSU
\]

Gulliksen concluded that speededness of a test can be determined by comparing the standard deviations of unattempted items (\(S_U\)) and items NOT marked correctly (\(S_E\) or \(S_{(W+U)}\)).\(^{42}\) (See Appendix A) Some of his conclusions were:

- Neither \(S_U\) nor \(S_W\) can be used to predict the magnitude of the other, thereby requiring any analysis to evaluate both \(S_U\) and \(S_W\);

---

\(^{39}\) Gulliksen, supra note 5, at 230-31.  
\(^{40}\) Rindler, supra note 3, at 263.  
\(^{41}\) Rindler, supra note 3, at 263; Gulliksen, supra note 5, at 232.  Note that the standard deviation is a measure of the scatter of values around an average.  The standard deviation, \(s\), of a sum, \((x+y)\), can be expressed as \(s_{(x+y)}^2 = s_x^2 + s_y^2 + 2rsxsy.\)  See Gulliksen, supra note 5, at 425.  The variable “\(r\)” is the correlation coefficient, which is a measure of how closely related two variables are.  Its value is always >= -1 (perfect negative correlation) and <= +1 (perfect positive correlation).  See Gulliksen, supra note 5, at 424.  
\(^{42}\) Rindler, supra note 3, at 263.
Only when W and U are perfectly correlated ($r_{WU} = -1$ or +1) AND one of these ratios, $S_U/S_{(W+U)}$ or $S_W/S_{(W+U)}$, is less than 0.1 can the equations below certify that a test is either primarily a speed test or primarily a power test:

**Speed Test** if $S_W/S_{(W+U)}$ is small, then: $1 + S_W/S_{(W+U)} > S_U/S_{(W+U)} > 1 - S_W/S_{(W+U)}$

**Power Test** if $S_U/S_{(W+U)}$ is small, then: $1 + S_U/S_{(W+U)} > S_W/S_{(W+U)} > 1 - S_U/S_{(W+U)}$

Unfortunately, conditions of perfect correlation and the specified ratios < 0.1 are extremely limiting. So, the model would have widespread utility only in rare circumstances. However, as applied by Gulliksen, the model has virtually no utility, due to his poor assumptions and a data classification problem that plagues the LSAT today. Gulliksen simply ignored rapid-guessing behavior, never classified skipped questions as unreached, did not differentiate between skipped questions that were considered and those that were not, etc. This meant that his estimates of W and U – as well as those of all researchers who adopted his model unchanged – were likely to be incorrect and undercounted. Even though Rindler altered definitions of W and U to compensate for some of this, she noted that it is impossible to know whether a given question went *unanswered* after it was fully considered or after time pressure convinced the test-taker to omit it; similarly, it is not possible to determine if an *answered* question was completed by the test-taker after considering the substance of the question or merely by rapid-guessing, due to time pressure.

---

43 The equations above, supplied by Rindler, did factor back into the model the questions skipped due to time pressure alone, considering them not reached – as with those left unanswered at the end of the test – and, therefore, part of U. She also placed questions considered, but not answered, back into the model by considering them part of the questions NOT correctly answered, W. However, she could not repair the significant data classification matter of how to differentiate among all of the above and rapid-guessing behavior for a paper-and-pencil test administration. In the end, that classification problem proves insurmountable for a pencil-and-paper test.

Gulliksen’s model, though helpful theoretically, cannot be used to estimate the presence of speededness reliably, whether they are answered or unanswered questions at the end, middle or beginning of any pencil-and-paper administered LSAT test, due to data collection and classification issues alone. It cannot be used in a realistic setting to calculate reliability estimates for today’s LSAT. It, therefore, cannot be used to estimate the degree of speededness present in an existing LSAT test or the impact of whatever speededness may be present.\footnote{This follows logically – if presence cannot be reliably detected, due to data classification errors, then neither can the degree of presence nor the implications of the presence. Gulliksen did derive reliability boundaries for speededness his model determined was present in a test, but the data classification issue clearly trumps these statistics and, in any event, his reliability calculations were rife with limiting assumptions. \textit{See} \textsc{Gulliksen}, supra note 5, at 231-243.} It can, however, be extremely useful when testing theoretical values and concepts, as well as when applied to computer-based testing. In those cases, Rindler’s notational corrections and accurate rapid-guessing behavior detection must be incorporated.

2.3.2 The Educational Testing Service’s (ETS’s) floor/ceiling model and LSAC’s per cent model
ETS accepts the underlying premises of Gulliksen’s model, but bypasses the difficulties in interpreting the quotients $S_U/S_{(W+U)}$ and $S_W/S_{(W+U)}$ by simply decreeing that a test is, effectively, a power test:

\begin{center}
\textbf{if at least 80\% of the examinees reach the last item and if everyone reaches at least 75\% of the items}. \footnote{Lu & Sireci, \textit{supra} note 5, at 33; Rindler, \textit{supra} note 3, at 263.}
\end{center}

There is evidence that LSAC has adopted a similar rule for the LSAT, that a test is decreed as a power test, if 10\% (or, possibly 15\%) of all test-takers reach the last item.\footnote{\textit{See} \textsc{Schnipke \\& Scrams}, \textit{supra} note 11, at 3. They note that: Other researchers switched from dichotomous decisions (power vs. speeded) to indices of speededness that are based on the percentage of test takers who do not reach all items.\footnote{\textit{This is the current \textbf{measure of speededness employed by LSAC with regard to the LSAT}. These indices are plagued ...}}
edicts have been widely criticized as being arbitrary.\textsuperscript{48} There appears to be no credible scientific research or evidence that these bounds permit a mixed speed and power test to function as though it were a power test.\textsuperscript{49} Worse, they suffer from the same data classification challenges as Gulliksen’s model. How does one attempting to apply this rule determine who has completed the test or reached 75% of it? Of course, for the LSAT, administered with pencil and paper and lacking response-time data recorded for each question, there is no consistent, accurate or practical method of determining whether a given question went \textit{unanswered} after it was fully considered or after time pressure convinced the test-taker to omit it; similarly, it is not possible to determine if an \textit{answered} question was completed by the test-taker after considering the substance of the question or merely by rapid-guessing, due to time pressure.\textsuperscript{50}

Also troubling and unjustifiable is that the rules amount to redefinitions of speededness, taking a generally continuous measure and transforming it into a binary that distorts any concept of the real definition. Speededness clearly occurs if any one of the 20% ignored by the ETS rule or the 10% ignored by the LSAT rule do not finish. Where Gulliksen provided a switch that was based on characteristics of power (and that was somewhat arbitrary as well), it

\textsuperscript{48} \textit{MEASURED PROGRESS, MAINE HIGH SCHOOL ASSESSMENT MECAS PART II: 2011-12 TECHNICAL REPORT 81 (2011-12), available at http://www.measuredprogress.org (last visited Dec. 15, 2013); Rindler, supra note 3, at 264. See also Lu & Sireci, supra note 5, at 33 (“this rule has been claimed to be an ‘arbitrary criterion’ and is ‘not rigidly applied’ (CEEB, 1984, p. 36).”).}

\textsuperscript{49} Rindler, supra note 3, at 263. She states: “The difficulty of interpreting the ratios proposed by Gulliksen presumably led the Educational Testing Service (ETS) to adopt a "rule-of-thumb" approach to assessing speed.”

\textsuperscript{50} See Rindler, supra note 3, at 263-64.
appeared to be narrowly tailored to meet the lower limit of the power/mixed test border. The ETS and LSAT rules make no such claim and, by their nature, ignore Gulliksen’s objective and attentiveness altogether. At the time the ETS rule was created, the educational testing community was fully aware that the SAT, LSAT and other standardized tests were timed such that African-Americans, Latinos and others could not fully consider all questions, at disproportionate rates. One LSAC study demonstrated that between 10% and 15% of African-Americans did not complete an LSAC test administration. The ETS and LSAT rules would ignore such facts and boldly declare that such tests were, nonetheless, fully acceptable tests of ability.

Perhaps worst of all is that such arbitrary definitions of speededness do not even specify that tests which fall short of the standard are invalid and unreliable tests of ability. These rules merely move the scientific goalposts by 10 or 20 per cent, freeing the test-makers of any burden to demonstrate that their tests really measure ability. Then, these rules never exact a penalty on test-makers for creating a self-declared speed instrument (if it meets the bounds of the respective rules) masquerading as a power test. What impacts on the test-taker, test validity or test reliability occur when the rules are not met? These edicts are unscientific, arbitrary mechanisms that grant a test-maker license to injure test-takers, particularly people of color, without consequence. Because they alter the very definition of speededness, they are

---

51 Id. at 262. Rindler states:
In an era which finds test constructors in a defensive stance with the public, claims that time limits unfairly penalize certain examinees have had their impact on attention to speededness. Of most interest have been assertions that time limits have more deleterious effects on blacks than whites, thus constituting a source of culture bias in testing.

52 Wightman & Muller, supra note 47, at 6.
53 Id. at 263-64.
54 Id.
incapable of measuring speededness and differentiating between power and speed tests. Any researcher or entity using these rules has ventured onto purely speculative ground.

Later, we show how such redefinitions dramatically distort the relationship between speededness for an individual and that of the test itself, a potentially devastating error among potentially large numbers of test-takers.

### 2.3.3 All other methods of identifying speededness are fatally flawed for pencil-and-paper exams

ETS researcher Donlon proposed a measure of speededness which extrapolates from the position in the test of the last question answered to estimate how long that test-taker would have required to complete the entire exam. These “time-needed estimates” form a distribution from which the administrator can identify a time limit under which a designated percentage of test-takers completed the exam. Presumably, this time limit will ensure certain power test characteristics. The test is considered a power test if the actual time limit falls within a certain confidence level of the calculated power test’s time limit. Unfortunately, this approach is not practical because it assumes that the level of question difficulty is uniform across the entire test; use of the last question answered ignores rapid guessing behavior at the end of exams and omissions earlier in the exam, due to time pressures; and it fails to calculate the penalty the time limit imposes – that is, the impact of the very speededness present.

---

55 Id. at 264.
56 Id.
57 Id.
58 Id.
59 Id.
Stafford proposed an assessment of speed measured by the portion of incorrect answers which are simply unanswered questions.\textsuperscript{60} He formed a “speededness quotient” for both the individual test-taker and, by summing across each test-taker, the entire test.\textsuperscript{61} Stafford expresses speededness in terms of frequencies, while Gilliksen expressed speededness as fractions of standard deviation.\textsuperscript{62} Unfortunately for Stafford, Gilliksen demonstrated that the proper mathematical relationships between the various factors is represented through the proportion of variance (or standard deviation), not linear frequencies.\textsuperscript{63} Thus, Stafford’s speededness quotient fails in theory and practice.\textsuperscript{64}

The single-test administration attempts to measure speededness all treat speededness as a function of the number of items a test-taker does not answer.\textsuperscript{65} This requires three assumptions which are rarely, if ever met.\textsuperscript{66}

\begin{enumerate}
\item all of the unattempted items are found at the end of the test;
\item all consecutive omissions that appear at the end of the test are unattempted (rather than considered and omitted) items;
\item unattempted items are the only elements of test performance that reflect speededness. \textit{Id.}
\end{enumerate}

These assumptions are poor and help to invalidate the measuring approaches.\textsuperscript{67}

Multiple or parallel test administrations provide a means for test-takers to compare their answers against a second or baseline set of answers.\textsuperscript{68} The second test is frequently

\textsuperscript{60} Id.
\textsuperscript{61} Id.
\textsuperscript{62} Id., at 265.
\textsuperscript{63} Id.
\textsuperscript{64} Id.
\textsuperscript{65} Id.
\textsuperscript{66} Id.
\textsuperscript{67} Id.
\textsuperscript{68} Id., at 266.
untimed and could be identical to the first. A test is considered a power test if additional time did not alter the before/after performance of each test-taker or, alternatively, the test-takers’ before/after rank among all test-takers for that administration. The measures here do not depend upon identification of unattempted items or even defining unattempted items.

Unfortunately, the administration of two tests is complicated and generally not practical.

Another approach to measuring speededness is to administer one test, but have an initial timed portion and a second untimed portion. The second portion may or may not permit the test-taker to go back to the timed portion and complete it fully. Three scores are computed: the number of correct answers in the timed portion, the number of correct answers in the untimed portion; and the total amount of time required to complete the exam. The administrator then analyzes the data with correlational or regression techniques. However, there are spurious correlations when scoring the same test multiple times, particularly when a test-taker is permitted to return to the timed portion after moving to the untimed portion. As well, calculating correlations among two different tests requires that they share sufficient characteristics or that estimates of expected correlation be calculated.

Rindler concluded:

[A]s currently formularized, assessments based on single test administrations are inadequate. Not only is no elegant single-administration based index of speed available,
but also, less refined methods depend on assumptions which are typically violated. Perhaps the most significant source of these violations is in the miscategorization of responses—items skipped before full consideration being treated as considered but omitted and series of considered and omitted responses at a test's end being treated as not reached.

Nevertheless, by their very nature, single-test administrations can yield no information on the extent to which time limits have effected overall pacing and potentially consequent response accuracy. Thus, it is most important that research be initiated in which the relationship of single administration based indices are compared with those derived from multiple administration methods which depend on fewer and less restrictive assumptions.

While Rindler reached her conclusions in 1979, it is abundantly clear that virtually nothing has changed in thirty years. Lu & Sireci conclude:

The literature is thin with respect to the relative effectiveness of different speededness evaluation methods, and so much more research in this area is needed. However, it should be stressed that testing agencies must employ some method for evaluating the time limits they impose on examinees. The specific amount of testing time allotted should be defended with respect to the interpretations that are likely to be derived from test scores.

In addition to detecting speededness, it is also important to remove its effects when it interferes with the construct measured by a test.

It is clear, then, that current models of speededness are not useful in practice, due to errors in the classification of responses implicit in pencil-and-paper testing. The extent to which the poor assumptions of the creators of the models infect the respective models is the extent to which the models themselves are not useful, in any regard.

2.3.4 Computer-Based testing
The data classification challenges described above largely evaporate when pencil and paper are replaced with computer-based exams. The item-classification issues disappear.

---

80Lu & Sireci, supra note 5, at 35.
because timing data is available for each question posed to each test-taker.\textsuperscript{81} Here is one description of how such item-response data can be exploited to arrive at far more accurate indicators of speededness:\textsuperscript{82}

Schnipke (1995) examined the item response times on an Analytical section from the Graduate Record Examinations (GRE) computer-based test. For each examinee, she graphed the standardized natural logarithm of response time of each item. The natural logarithm was used to normalize the distribution. She observed that a typical speeded examinee who switched from “solution behavior” (response after careful consideration) to “rapid-guessing behavior” usually responded at a slower rate than most examinees at the beginning of the test. After some items, the examinee would suddenly start responding much faster, accompanied by a noticeable drop of accuracy. To check for speeded items, she examined response time distribution for each item together with proportion of examinees giving correct responses at each response time level. She found that items appearing later in the test have more combinations of little response time and inaccurate responses.

She also performed a series of ANOVAs in which the response time was predicted from response accuracy and item characteristics. It was found that the variation in items was less predictive of response times on the second half of the test, implying that some examinees were not fully considering the items.

Using the Item Response Theory, which estimates the probability of a correct response to a question (or item) through models, and exploiting computer-based testing to examine how long a test-taker spends on each question,\textsuperscript{83} LSAC researchers have demonstrated that it is possible to estimate which questions were fully considered and which were not, for any variety of reasons.\textsuperscript{84} These researchers directly and unequivocally warn that LSAC’s current means of measuring speededness is likely to be inaccurate, while deriving item-response data from computer-based testing would give highly accurate speededness calculations:

\textsuperscript{81} Id. at 34.
\textsuperscript{82} Id. at 34.
\textsuperscript{84} Schnipke and Scrams, supra note 11, at 15.
Traditional speededness indices depend on “unreached” items, but when test takers engage in rapid-guessing behavior, the items are not unreached. More recent approaches to speededness depend on the drop in accuracy at the end of the test ... The latter approach is better because it incorporates guessing, but a more direct approach would focus on the amount of time spent on each item.\(^{85}\)

These LSAC researchers used a computerized Graduate Record Examinations (GRE) General Test administration as their problem domain.\(^{86}\) They analyzed two item types, logical reasoning and analytical reasoning, which they considered similar to the corresponding LSAT types.\(^{87}\) We extrapolate from their results to the LSAT, on a largely theoretical basis.\(^{88}\) Based on their conclusions, the LSAT could be speeded for as many as 15 per cent of the test-takers, using the traditional definition,\(^{89}\) and as many as 50 per cent, using the newer definition of Schnipke and Scrams, based on item-response times.\(^{90}\)

### 2.4 The effects of speededness

#### 2.4.1 A test is speeded if and only individuals find it speeded: Reasserting the real definition of speededness

Being able to detect speededness with little guesswork or extrapolation would seem to be an extraordinary advantage of computer-based testing over pencil-and-paper testing. However, as noted above, there is a dual use of the word “speeded” by test-makers and researchers, namely that a test can be unspeeded, even though a significant portion of the test-takers find it speeded. An example of a formulation of the ETS, LSAC, etc. definition of the term speeded, when applied to a test, is:

\(^{85}\) Id.  
\(^{86}\) Id. at 2.  
\(^{87}\) Id.  
\(^{88}\) Id.  
\(^{89}\) Id. at 16.  
\(^{90}\) Id.
In other words, “a test is speeded when some portion of the test-taking population does not have sufficient time to attempt every item in the test within the allocated time” (Bejar, 1985, p. 1). (emphasis added)

This “some portion” aspect of that definition differs from the Gulliksen’s theory, which is that a test is speeded if it is speeded for any portion of the test-taking population. The ETS-like definition seems poised to allow for arbitrary cutoffs that ignore the injury to those below that demarcation line for speededness, wherever that line may lie. Similarly, LSAC researchers Wightman and David G. Muller identified an LSAC cutoff standard:

In general, a test section is not considered speeded if over 90 percent complete it.91 (emphasis added)

The subtle language shift reflected by Wightman and Muller similarly confuses and contorts the analysis of speededness. Such notions transform detection of speededness into an exploration of a safe harbor for test-makers. This percentage definition, effectively, redefines speededness into a binary yes/no – unrelated to the underlying concept. It shields test-makers from criticism, ignores potentially dramatic injury to test-takers, and denies them remedy. Test-Takers and plaintiffs in legal actions against LSAC must reclaim the proper use of the word “speeded” or they will start any lawsuit at an immediate deficit.

In this section, we demonstrate that speededness for an individual is the same principle as speededness for a test and that any definition for test speededness necessarily is based upon speededness for an individual test-taker. The ETS and similar definitions of test speededness must be seen as disconnected from the underlying concept and improper.

Attempts to examine speededness within a framework that involves simple percentages of persons appearing to complete their respective tests is farcical and is refuted by educational

---

91 Wightman & Muller, supra note 47, at 6.
testing theory. The contribution of individual speededness to the aggregate test statistics varies non-linearly through standard deviations. The 80% and 90% rules do not approximate such calculations.

Beyond these concerns with percentage rules is a foundational one. When researchers or test-makers create definitions that intentionally discount large portions of the test-takers (10 per cent, 20 per cent etc.), and begin counting speededness anew, the line is so stark and the effort so pronounced that there is an expectation that the bold standard somehow imparts a seal of approval on those within its bounds and a claim of invalidity and/or unreliability on those outside its boundary. However, nothing in the literature suggests that any of the cutoffs identified thus far carry with them a series of well-researched conclusions about how speededness, so identified, so dramatically impacts the test-takers and/or the test that the term should suddenly apply. Fortunately, Gulliksen’s equations do make cautious statements about when a speeded test, as measured via the standard deviation, is incapable of being a measure of ability. The suggestion, then, is that the percentage standards serve no scientific or measurement purpose whatsoever, but might better serve a public relations or liability purpose.

2.4.1.1 Reminder of what researchers state speededness is

Researchers have typically discussed speededness as a continuum of impacts on test-takers resulting from applying time constraints to a power test:

When time’s limitation affects examinees' scores, a test becomes a partially speeded instrument; that is, one in which scores are determined by the number of items attempted (speed of response) as well as the accuracy of the supplied answers (power
of response). The larger the proportion of items examinees lack time to attempt, the greater is the speededness of the particular test.\textsuperscript{92} (emphasis added)

Gulliksen noted this phenomenon as present if even a single test-taker were unable to complete even one question. He stated:\textsuperscript{93}

\[ U \text{ [number of unattempted questions] will be zero for each person ... } \]
\[ \text{To the extent that these conditions are approximated, the test approaches a power test.} \]
\[ \text{...} \]
\[ \text{In a pure power test, all subjects will finish, the variance of } u \text{ will be zero ...} \]
\[ \text{...} \]
\[ \text{if the number correct is used, both speed and accuracy enter into the score. By using both these scores, we can determine the relative reliability of speed alone and of speed together with accuracy.} \textsuperscript{94} \text{(emphasis added)} \]

Gulliksen is very clear that in a pure power test, \( U = 0 \) for each test-taker. If one test-taker does not complete an exam due to time constraints, the test is, by definition NOT a pure power test. If scoring depends on factors other than the number unattempted, then neither is the test a pure speed test. Hence, when \( U \geq 1 \text{ for even 1 test-taker} \) in a test where the number correct matters, there is a mixed speed and power test. Gulliksen’s model, in theory (though highly constrained or impractical), could then be used to gauge the relative contributions of speed and power to see if the test “approached” a speed or “approached” a power test.

Note that Gulliksen did not use the term unspeeded.\textsuperscript{95} Nor did Gulliksen use the term speeded in his discussion, but he used it as a universal replacement for “speed” in his index

\textsuperscript{92} Rindler, supra note 3, at 261.
\textsuperscript{93} GULLIKSEN, supra note 5, at 231.
\textsuperscript{94} Id. at 238.
\textsuperscript{95} Id. at 484. This is corroborated by the index entries:
  United States Navy, 209, 265, 267, 273, 304
  United States War Department, 418
  University of Chicago, The, 2, 40(n), 204, 218(p), 229(p), 272, 286, 418
  Validity, effect of explicit selection on, 137-138, 142, 148-150,156
The extent to which Gulliksen spoke for the educational testing community is the extent to which it is categorical that speededness for an individual and speededness for a test are precisely the same concept. Did even one test-taker have insufficient time to answer one question? If that is so, then the test should not be characterized as unspeeded. Gulliksen noted and explained, however, that some levels of speededness do not deprive a test of its latent power characteristics and he endeavored to identify such a threshold.

2.4.1.2 Deriving test speededness from individual performance

Gulliksen derived a model of the texture of speededness – for both individuals and a specified test – from a mathematical perspective. Though he incorporated data elements that are not measurable outside computer-based testing and Item Response Theory, and perhaps not even then, Gulliksen’s basic concepts are still considered valid. We can return to his work, briefly, to see precisely how misleading the concept of a test being “unspeeded,” due to an ETS-like proportional threshold, really is.

In Appendix B, we demonstrate that individual speededness contributes to test speededness on the order of \( \sqrt{\sum (U_i - M_U)^2 / T} \), where each test-taker is identified by a different index \( i \), \( T \) is the total number of test-takers, each test-taker’s number of unanswered questions

---

96 Id. at 483. Examples:
- Spearman’s correction for attenuation, 101-104, 105
- Speeded test, correction of score for wrong answers, 251-252, 260
  - definition of, 230-233, 241
  - effect of guessing on score, 246-251
  - error of measurement for, 233-236, 242
  - item parameters affected by, 385-386
- Speeded test, odd-even reliability for, 236-238, 242
  - variance change in, 231-233, 241-242
- Standard deviation (see also Variance)

97 Rindler, supra note 3, at 262.
is \( U \), and the mean of \( U \) is denoted by \( M_U \). We use this fact to illustrate how an ETS-like percentage cutoff (80% or 90%) masks substantial individual speededness and bears no relationship to underlying speededness concepts. Through an analysis of Gulliksen’s equations and three illustrations using sample LSAT test administration data points, we have been able to demonstrate the following:

- A percentage cutoff for speededness, such as that used by ETS and LSAT, bears little resemblance to the models of speededness that define power and speed tests because the cutoffs indicate nothing about and require nothing of standard deviation and correlation, through which speededness is expressed (we chose \( S \) and \( r \) in the illustrations for demonstrative reasons, but nothing in the ETS or LSAT rules constrained those choices);

- In an illustration of an LSAT exam in which 90% of the test-takers experienced no speededness, while 10% did not complete 15% of the test, due to time constraints, we held ability roughly equal for the portions of the test that both groups completed (\( W \) values were approximately equal). In that scenario, the test appeared to have the characteristics of a speed test and bore little resemblance to those of a power test. Fundamentally, this indicates that students of equal ability can be scored differentially — **exclusively due to speededness**. LSAC characterizes a 90% cutoff as “slight” speededness, while the ETS model suggests this percentage is well within the bounds of a power test — utterly devoid of speededness. This illustration demonstrates that both models can be incorrect, given modest and practical assumptions;

- In an illustration of an LSAT test that is a pure power test for 80% of the test-takers and maximally speeded, per the ETS standard, for the remaining 20% (they did not reach 25% of the test), the test has no characteristics that approach that of a power test. This actually means that the ETS standard predicted speededness correctly, in this instance. However, it also means that the level of speededness was so great at the 80% cutoff that the test could no longer be considered a test of ability. This suggests that, in some circumstances, the ETS standard might not mark the beginning of speededness in a harmless, academic characterization, but the pronounced effects of speededness;

- Individual test-takers for whom the test was speeded can overwhelm the speededness characteristics of a test through their non-linear contributions to relevant standard deviations. Speeded test-takers comprising only 10% of the test-taking population can contribute to the variance of \( U \), \( W \) and \( (U+W) \) by an order of magnitude greater than the combined effect of all test-takers who completed the test. This indicates the intensity of the impact of speededness and helps to explain why speededness for a few individuals must be viewed with concern. It also explains why an arbitrary cutoff at 80% or 90% bears little relationship to overall or individual speededness.
In effect, it is easy to show that speededness of a test is purely a function of speededness of individual test-takers and that LSAT/ETS-like percentage cutoffs do not indicate speededness or lack thereof. Such rules cannot consistently assist in the determination of speededness, irrespective of their claims.

2.4.2 The impacts of speededness on reliability and validity
Since the percentage rules of thumb (90% cutoff, 80% cutoff) clearly do not permit a categorization of the effects of speededness for those tests that are on either side of their arbitrary speededness lines of demarcation, we must identify these effects by other means. Lu & Sireci surveyed possible impacts of speededness.98 However, there does not appear to be a catalogue of impacts indexed to level of speededness (and there is no gauge for level of speededness in wide use, at any rate). Ultimately, such evidence will be essential to demonstrating whether the LSAT is or was ever a scientifically sound instrument.

For a standardized test to be useful, it must be both reliable and valid.99 Reliability is a measurement of how accurate a test score is and there are many types of reliabilities.100 The reliability coefficient measures the error component in a test score and is viewed as correlation coefficients are (values between -1 and +1, values > 0.7 are considered acceptable).101 As well:

[I]nternal consistency ... is probably the most often quoted reliability. It is a measure of whether each question in a test measures the same thing. There are various ways of calculating this figure ... [including] Cronbach’s alpha [and older measures] ... such as Kuder-Richardson KR20 and KR21 and Rulon. [These statistics, sometimes called split-

---

98 Lu & Sireci, supra note 5, at 29-37.
99 JAMES BOYLE & STEPHEN FISHER, Unit 3: The Importance of Reliability and Validity, EDUCATIONAL TESTING: A COMPETENCE-BASED APPROACH TEXT FOR THE BRITISH PSYCHOLOGICAL SOCIETY’S CERTIFICATE OF COMPETENCE IN EDUCATIONAL TESTING 63 (February 11, 2008).
100 Id. at 65.
101 Id.
half reliability,] split up the test in various ways and correlate combinations of questions together, for example, odd with even, first half of the test with second half.  

Gulliksen examined split-half reliability in the presence of speededness. He noted that for a mixed speed and power test (one in which \( S_U \) is negligible), split-half reliabilities are an “unsafe basis for estimating the test reliability.” He determined that reliability coefficients must be calculated by taking correlations between the subject test and a parallel test, with the same test-takers taking both tests. Lu & Sireci note that Gulliksen found the underlying issue to be that reliability estimates are “spuriously high” in speeded environment. Critical for our purposes is that speededness must be considered by LSAC researchers who use split-half reliabilities to characterize the reliability estimates for the LSAT. Failure to assess the presence of speededness in the LSAT test being analyzed could produce reliability results that are “spuriously high.”

LSAC produces ten statistical reports on a regular basis characterizing various aspects of each LSAT administration. They describe these reports as:

The LSAC Psychometric Research staff regularly monitors various psychometric and statistical properties of the Law School Admission Test (LSAT). LSAT Technical Reports document the results of these analyses on a regularly scheduled basis. The table below describes the technical report topics with their respective replication cycles.

LSAC identifies the names and frequency of these reports, as described in Table 1 below.

102 Id.  
103 GULLIKSEN, supra note 5, at 233-35.  
104 Id. at 235.  
105 Id.  
106 Lu & Sireci, supra note 5, at 30.  
108 Id.
It is clear that LSAC does not track or even account for speededness or reliability as a matter of standard practice. For instance, in the current performance report, *LSAT Performance With Regional, Gender, and Racial/Ethnic Breakdowns: 2005–2006 Through 2011–2012 Testing Years*, LSAC does not include the terms “speeded,” “speededness” or “reliability” even once.\(^{109}\) Neither are these terms included in the current *National Summaries of Correlation Studies* report, called *Predictive Validity of the LSAT: A National Summary of the 2009 and 2010 LSAT Correlation Studies*.\(^{110}\) A review of selected LSAC technical reports and LSAC’s research report series indicates that these terms are absent as a rule. LSAC is clearly not measuring reliability with respect to speededness, but neither is it measuring reliability, on a discernible basis,

---


independent of speededness. This has significant implications for defenders of the LSAT in a

law suit.

It is possible to have a reliable test that measures nothing of interest.111 The concept of validity is used to determine whether the test measures what it states that it measures.112

There are four types of validity:113

- Face validity – qualitative measure that addresses whether the test “looks as if it measures what it says it measures,” affects test’s acceptability from test-taker’s view;
- Content validity – degree to which test questions are representative of what the test intends to measure, for example an arithmetic test should include division;
- Construct validity – extent that a test measures the underlying concept it intends to measure, usually evaluated though correlations with a trusted benchmark test;
- Criterion-related validity – there are two types, concurrent and predictive, each of which attempts to identify correlations with a criterion associated with the purpose of test, but which is difficult to measure directly:
  - Concurrent validity – the criterion is being measured at the time the test is taken;
  - Predictive validity – the criterion is to be measured at a future time, such as first-year GPA in law school (FYA).

Quoting the American Educational Research Association (AERA), Lu & Sireci provided a broader definition of validity:114

“Validity is the most fundamental consideration in developing and evaluating tests. The concept refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9).

Lu & Sireci demonstrated that speededness impacts construct validity.115 A primary assumption of testing is that a score should accurately measure traits that the test is designed to measure and that differences between scores should accurately reflect differences in the degree to

111 BOYLE & FISHER, supra note 99, at 65.
112 Id.
113 Id. at 65-66.
114 Lu & Sireci, supra note 5, at 30-31.
115 Id. at 31.
which test-takers possess that trait.\textsuperscript{116} Speed introduced into a test of ability, as the LSAT is described, is a new trait for which no score is assigned and which influences the measurement of the examined traits in unpredictable ways.\textsuperscript{117} The result is that inferences about ability, derived from test scores, can no longer be reliably made.\textsuperscript{118}

Speededness can also affect content validity.\textsuperscript{119} Consider a test that is meant to measure specific abilities through the use of the last few questions on the test.\textsuperscript{120} However, the introduction of speededness does not permit all test-takers to reach those questions.\textsuperscript{121} Thus, the test under-represents the specified content.\textsuperscript{122} It can no longer measure those abilities and it is, hence, invalid as a measure of said abilities.\textsuperscript{123}

Speededness impacts all four types of validity by biasing a test differentially for race and other characteristics.\textsuperscript{124} African-Americans and Latinos have been found to require more time on SAT tests than their White counterparts of comparable abilities.\textsuperscript{125} Speededness would likely cause Whites to outperform Blacks and Latinos on such tests, but that result would be invalid by prior assessment of the abilities of the test-takers.\textsuperscript{126}

There have been studies that purport to demonstrate that speeded tests show comparatively poorer performance for African-Americans than others.\textsuperscript{127} Yet, Lu & Sireci seem

\begin{thebibliography}{99}
\bibitem{116} Id.
\bibitem{117} Id.
\bibitem{118} Id.
\bibitem{119} Id.
\bibitem{120} Id.
\bibitem{121} Id.
\bibitem{122} Id.
\bibitem{123} Id.
\bibitem{124} Id.
\bibitem{125} Id.
\bibitem{126} Id.
\bibitem{127} Id.
\end{thebibliography}
to accept, giving African-Americans more time does not help them close the gap.\textsuperscript{128} A closer look at the studies on which Lu & Sireci rely for their conclusion, however, reveals that there is no basis on which to conclude that African-Americans were given sufficient time to demonstrate their ability even when the extra few minutes were added.\textsuperscript{129} Perhaps they needed additional hours, for instance, because their White counterparts had purchased access to expensive LSAT preparation courses; had been drilling incessantly for three months; and had become unnaturally swift when answering LSAT questions. Research has shown that additional time can help persons with disabilities to increase their scores significantly.\textsuperscript{130} It is not unrealistic to conclude that sufficient time increments can help all test-takers.

LSAC’s various predictive reports capture predictive validity estimates using first-year law school GPA (FYA) as the criterion being predicted. Yet, these reports do not reference speededness. Even in the absence of addressing potential speededness, the correlations cited

\textsuperscript{128} Id.

\textsuperscript{129} See Franklin R. Evans & Richard R. Reilly, A Study of Speededness as a Source of Test Bias, 9 J. EDUC. MEAS. 123, 123-131 (1972). The so-called unspeeded exam simply required test-takers to answer 27 of the 35 Reading Comprehension test questions, while the so-called speeded test required that all 35 be answered. A comparison was made between the performance of both groups on the 27 questions only. The authors admitted that they were unaware of the methodological soundness of this arrangement with:

Even though the speeded and unspeeded tests contained the same 27 items, they should not be considered identical tests since the speeded version contained 8 additional items which could have affected the scores in some unknown way. Id. at 124-25.

In fact, Rindler criticized this type of comparison as wholly unscientific and utterly useless:

The technique of simply computing correlations between timed and untimed test scores or two halves of the same test is meaningless without some basis of comparison ideally, the expected correlation between two administrations of the materials under the same time limit. Rindler, supra note 3, at 268.

The study also assumed that African-Americans and Whites were being measured due to the location of testing centers. There is no definitive basis for this characterization whatsoever. Thus, the very subjects of the research were not individually identified racially. This study is often cited, but appears to have no scientific merit whatsoever.

\textsuperscript{130} Id.
by the various predictive validity studies are so weakly correlated as to be negligible and the correlations have been decreasing significantly over time, as shown in Table 2.\textsuperscript{131}

<table>
<thead>
<tr>
<th>LSAC Studies: National Summaries of Correlation Studies</th>
<th>LSAC Published Value of Correlation Coefficient ($r$)</th>
<th>% of Variance in FYA Explained by LSAT: Coefficient of Determination ($R^2$)</th>
<th>% of Variance in FYA NOT Explained by LSAT: Coefficient of Determination $(1 - R^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>0.40</td>
<td>16.00%</td>
<td>84.00%</td>
</tr>
<tr>
<td>1996</td>
<td>0.40</td>
<td>16.00%</td>
<td>84.00%</td>
</tr>
<tr>
<td>1997</td>
<td>0.39</td>
<td>15.21%</td>
<td>84.79%</td>
</tr>
<tr>
<td>1998</td>
<td>0.39</td>
<td>15.21%</td>
<td>84.79%</td>
</tr>
<tr>
<td>1999</td>
<td>0.39</td>
<td>15.21%</td>
<td>84.79%</td>
</tr>
<tr>
<td>2000</td>
<td>0.40</td>
<td>16.00%</td>
<td>84.00%</td>
</tr>
<tr>
<td>2001</td>
<td>0.40</td>
<td>16.00%</td>
<td>84.00%</td>
</tr>
<tr>
<td>2002</td>
<td>0.39</td>
<td>15.21%</td>
<td>84.79%</td>
</tr>
<tr>
<td>2003</td>
<td>0.37</td>
<td>13.69%</td>
<td>86.31%</td>
</tr>
<tr>
<td>2004</td>
<td>0.35</td>
<td>12.25%</td>
<td>87.75%</td>
</tr>
<tr>
<td>2005</td>
<td>0.34</td>
<td>11.56%</td>
<td>88.44%</td>
</tr>
<tr>
<td>2006</td>
<td>0.33</td>
<td>10.89%</td>
<td>89.11%</td>
</tr>
<tr>
<td>2007</td>
<td>0.32</td>
<td>10.24%</td>
<td>89.76%</td>
</tr>
<tr>
<td>2008</td>
<td>0.33</td>
<td>10.89%</td>
<td>89.11%</td>
</tr>
</tbody>
</table>

Notes:
1. $R^2$ explains the % of variation in FYA attributable to the LSAT.
2. $(1 - R^2)$ is the % of variation in FYA unexplained by the LSAT.

The validity statistics show that correlation coefficients are now on the order of 0.3, far less than the 0.7 considered acceptable throughout behavioral sciences\textsuperscript{132} and that approximately 90% of the variation in FYA is unexplained by the LSAT. A natural question is: What does this


\textsuperscript{132} Boyle & Fisher, supra note 99, at 65.
90% actually explain? Factoring in potential speededness calls these results further into question.

Speededness is likely affecting the very core of what the LSAT is marketed as measuring, turning a test with poor reliability and validity indicators into an even more dubious tool.

2.5 Conclusion

Speededness is a concept in educational testing that refers to the impact of time pressures on individuals who are taking tests that purport to be measures of ability. Under sufficient time pressure, test-takers are unable to consider each question fully. They skip questions in an attempt to answer more questions later in the test; they leave questions unanswered at the end; they engage in rapid-guessing behavior as the timed sessions end. A test that is sufficiently speeded has test-takers never reaching significant numbers of questions and suffering scoring degradations as a direct result. Because these tests fail to measure ability alone, their scoring cannot be interpreted as gauges solely of ability, but of a combination of both ability and speed of response.

Unfortunately, for the LSAT, which is given by pencil and paper, there is no data recorded or provided to test-takers and law schools of the time a test-taker needed to complete each question or section. This data would help to determine which questions were fully considered and which received rapid guesses, as the clock expired. Neither is there any record of the time a test-taker used to read and consider a question – without answering it, or a record of whether a test-taker simply skipped a question, without ever having read or considered it. These missing data points make it impossible to determine if any given LSAT test was speeded for one or all test-takers.
The result of LSAC’s failure to record and evaluate response times, and to classify each response properly, is that LSAC is now incapable of certifying that any of its tests have ever been unspeeded for any individual or for the tests themselves. LSAC has adopted various measures of speededness, based on the per cent of those who “reach” the end of a section or test. However, these approaches fail because there is no means of determining what “reach” means, in the absence of item response times. They also fail because researcher Gilliksen, as interpreted by Rindler, demonstrated that speededness varies not by linear percentage of unanswered questions, but by the standard deviations of (a) the number of questions which are both unanswered and unconsidered, and (b) the number of incorrect answers, which includes those fully considered, but unanswered. Hence, the 80 per cent, 90 per cent and other per cent lines of demarcation for speededness, decreed by ETS and LSAC, are not tied to the underlying theory of speededness and must, of necessity, produce incorrect results. Numerous models exist for measuring speededness, but they all suffer from the same data classification problem or theoretical shortcomings, particularly when item response times are unavailable.

There are also significant challenges to the reliability (accuracy) and validity (measurement consistency) of the LSAT that only become more problematic when speededness is taken into consideration.

Irrespective of whether and how test-takers can demonstrate the above points, it is clear that LSAC cannot demonstrate that the LSAT is a reliable and valid measure of ability. The LSAT, therefore, is vulnerable to legal challenges on evidentiary grounds.
3 Legal Action against LSAC and Other Parties

We can countenance a law suit, to be fought within the next six months, in which *the lack of scientific evidence* will win the jury or convince the defendant(s) to settle. The plaintiff will challenge a law school’s denial of her application for admission. That denial will be based, in part, on the applicant’s LSAT score(s). Below, we sketch the general contours of such a suit and explore its more pronounced hurdles.

3.1 Injury caused by the LSAT

The LSAT has consistently reported that African-Americans, Puerto Ricans, and other people of color are not to be considered as capable in law school as Whites and Asians, based upon their low LSAT scores. At precisely the same time, LSAC has evidence that the LSAT is speeded for those groups and there is evidence that the LSAT may be speeded for as many as half of all test-takers. We have shown, above, that LSAT scores cannot be considered indications of ability when speededness is pervasive. At worst, we have demonstrated that LSAC is *incapable* of proving that the LSAT functions as a power test. The damage caused by this instrument is wide-spread. That injury has accrued to:

- **Individuals** – Test-Takers were assigned a score that the test-taker deems to be demeaning; there is a denial of opportunity to attend the test-taker’s law school of choice; there are wasted testing fees, application fees and preparation resources;
- **Groups** – African-Americans and Puerto-Ricans, in particular, are denied entry to better law schools unjustly;
• Society – With law school barriers, there are invariably fewer lawyers of color; less justice for the clients they would have served; fewer role models; less diverse law schools;

• Justice System – Fewer lawyers means less diversity among judges.

There is significant injury caused by an exam that advertises itself to be a meritorious instrument, when, in fact, it is a discriminatory tool that prevents diversity within the criminal justice system.

3.2 Liability for injuries caused by the LSAT

LSAC’s actions have made it most susceptible to liability because it creates and administers the LSAT each year, so it is directly responsible for a test that unlawfully injures individuals and groups alike. Law schools were aware or should have been aware of all issues presented here – we have relied on public information. Law schools also have access to all LSAC studies and specific numbers for their own schools - because LSAC sends each law school such information free of charge.133 Being able to choose a compensatory model of admissions or rely on undergraduate GPA – and not the LSAT – were available to law schools. That they chose a path that they knew or should have known produces discriminatory outcomes suggests intentional discrimination on the part of law schools. The ABA oversees law school admissions and the use of the LSAT; they either know of should have known about its injuries. The DOE oversees the ABA’s relationship to law schools by contract/agency. US News defames law school applicants through the use of invalid, unreliable scores and creates a culture of

133 LSAC, Cautionary Policies Concerning the LSAT, supra note 15, at 1.
adherence to the LSAT – even though they must know or should know of the test’s continual injury.

### 3.3 Heart of the lawsuit

Core to the lawsuit will be that the defendants unjustifiably and rigidly relied upon the LSAT, which the plaintiff will describe as an unscientific and unproven admissions test. The plaintiff will further note that evidence of the systemic flaws of this test were available and in the public domain for decades. The plaintiff will argue that the unscientific nature of the test stems from its application of unnecessarily strict time limits to a test that is advertised and widely believed to be a test of ability. The plaintiff will note, almost parenthetically, that if LSAC considered the time that a test-taker required to complete any portion of the test to be important, LSAC would report question-, section-, or test-completion times, possibly providing percentile rankings of each, as it does with current LSAT number-correct scores. Clearly, then, LSAC considers this to be a test of ability, whose time constraints are, at best, matters of administrative convenience.

Plaintiff will continue that the imposition of harsh, narrow time limits has forced many, if not most, test-takers to rush through the exam, failing to consider each item fully, frequently skipping questions to get to the next one, and frequently engaging in rapid guessing as the clock expired in each section. The plaintiff will note that testing research has proven that a test of ability, also called a power test, cannot be timed in a manner that deprives test-takers of the opportunity to consider each and every question fully. When that occurs, the plaintiff will state, the test ceases to be a test of ability and becomes, at least in part, a test of speed – how fast can one rush through the exam.
Plaintiff will charge that, as a result of LSAC’s restrictive time limits, vast numbers of people, perhaps half of all test-takers, likely the average African-American and Puerto Rican test-takers, did, in fact, have too little time to consider every question, therefore leaving many unanswered, rapidly guessing, and leading to test scores that were lower than their respective abilities and test-taking skills merited.

3.4 Possible Counts I, II and III of lawsuit

Count I could be a breach of contract claim. Plaintiff would allege that the act of paying a testing fee to LSAC for a specific LSAT administration, and LSAC allowing entry into a testing center to take the test, and later scoring said test, a contract had been formed that required LSAT to provide a scientifically valid, reliable measure of ability. There was also a contract formed when a test-taker used a fee waiver to gain entry to an LSAT administration, because the test-taker detrimentally relied on LSAC properly scoring, ensuring the validity and reliability of the LSAT, and reporting accurate test results. There was a breach when LSAT arbitrarily administered a test with capricious time limits and arbitrary scoring, then reported it, as valid and reliable indicia of plaintiff’s ability, to all law schools where plaintiff sought admission.

Count II could be a negligence claim. Plaintiff would argue that LSAC had a duty to exercise reasonable care in the administration of a test that was demonstrably valid scientifically and non-prejudicial to any test-taker. LSAC also had a duty to be truthful when reporting the abilities of the plaintiff to any entity. Law schools had a duty to exercise reasonable care in the processing of applications. A breach occurred in LSAC’s negligent development and administration of plaintiff’s test, negligent scoring of plaintiff’s test, negligent reporting of plaintiff’s score, and negligent supervision of LSAC employees. A breach occurred
in the negligent processing of applications, through the use of a test that the schools were aware could not be demonstrated to be valid and which produced biased results, as well as in negligent supervision law school employees. A breach occurred when LSAC defamed plaintiff by falsely reporting that it had evidence that plaintiff was not as prepared to be successful in law school as those ultimately accepted by the law school. The injury was Intentional Infliction of Emotional Distress (IID) and Negligent Infliction of Emotional Distress (NIED) in being unfairly denied admission, and being defamed.

Count III could be a Civil Rights claim. Plaintiff would assert that law schools, the ABA, LSAC, DOE, and US News magazine intentionally discriminated against plaintiff. Civil Rights precedent requires that plaintiff proves intentional discrimination. Plaintiff will rely on City of Canton v. Harris, 489 U.S. 378 (1989) (holding that where a Title 42 U.S.C. “§ 1983 plaintiff can establish that the facts available to city policymakers put them on actual or constructive notice that a policy is substantially certain to result in the violation of the constitutional rights of their citizens, ... [only] then can it be said that the municipality has made ‘a deliberate choice to follow a course of action . . . from among various alternatives.’”). Plaintiff will point out that there were non-discriminatory alternatives to the current LSAT, including an untimed LSAT, use of undergraduate GPA (UGPA), and compensatory admissions models. Plaintiff will note that research studies disparaging the predictive capability of UGPA correlate it arbitrarily, and counter to public interest, with FYA and not with second-year GPA, third-year GPA or bar passage rate.
3.5  **Daubert hearing – Admission of LSAC’s scientific evidence is unlikely**

As fully documented above, LSAC will be forced to prove the validity and reliability of the LSAT, including that the LSAT suffers no ill effects from speededness. This will be impossible, in part, because the LSAT’s pencil-and-paper mode of administration has never permitted the recordation of question response times. While beyond the scope of this paper to complete a strategy for the required evidentiary hearings, it is helpful to sketch an outline of the standards by which the battle for the admission of LSAC’s evidence will likely be judged.

In federal court and most state courts, the reliability and sufficiency of scientific evidence must be established in a Daubert hearing before such evidence can be admitted or an expert witness can testify.\(^{134}\) In *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993), the Supreme Court abandoned court deference to the broader scientific community by introducing a four-part test for ensuring the reliability of expert testimony in federal courts.\(^{135}\)

The factors a judge must consider before admitting scientific evidence include:

- Whether it can be (and has been tested); *Daubert*, 509 U.S. at 643.
- Whether the theory or technique has been subjected to peer review and publication; *Id.* at 644.
- Known or potential rate of error; *Id*.
- General acceptance. *Id*.

The idea is to ensure that the evidence is scientifically valid. *Id.* To this end, any attempt by a law school to introduce LSAT test scores into a law school admissions lawsuit must be met with strenuous objections on the grounds of relevance, due to the fact that


\(^{135}\) *Id.* at 637-38.
scientifically invalid evidence is not relevant. Id. Similarly, a torts case alleging negligence, IID, NEID or defamation for the negligent and/or intentional use and disclosure of unreliable and invalid test scores, falsely promoted as scientifically valid, would require LSAC and law schools to defend their actions by first demonstrating that the LSAT meets the Daubert criteria. A breach of contract case would allege that the plaintiff was not provided with a demonstrably scientifically valid test when the plaintiff took the LSAT. LSAC would argue that they did provide plaintiff with such test and would proceed to enter LSAT evidence into the record. Plaintiff would object on the grounds of relevance, as above. A Civil Rights claim would be defended with arguments challenging that there ever was discrimination and/or that any discrimination was intentional. The defendants would invariably attempt to state the venerable qualities of the LSAT as a “tried-and-true,” well-researched pillar of the legal education system, to which the plaintiff would object on the grounds of citing facts not in evidence and, as above, relevance.

The Daubert factors do not weigh in LSAC’s favor because:

- Testing – Speededness has been documented, but LSAC does not repeatedly test the LSAT, pursuant to its technical reports, to ensure and demonstrate that its tests are, essentially, power tests. In fact, LSAC will fail to demonstrate that its tests are power tests as long as they remain pencil-and-paper tests;

- Subjected to peer review and publication – LSAC has never publicized its ongoing efforts to ensure that its tests are not speeded; LSAC does not publish reliability analyses for each of its tests;
• Known or potential rate of error – LSAC cannot demonstrate its rate of error associated with speededness or reliability. LSAC’s data collection and data classification issues will prevent this demonstration as long as the LSAT remains a pencil-and-paper test;

• General acceptance – LSAC will not be able to find a sufficient body of scientific evidence to demonstrate that its failure to report on speededness or its failure to record question response time are generally accepted practices.

Any legal attack on the LSAT can and must exploit evidentiary problems faced by proponents and willing consumers of the LSAT.

4 Conclusion
This paper explored the failures of LSAC to research and protect against the impacts of unnecessarily short time limits imposed by LSAC on LSAT test-takers. We examined the science of educational testing to see, first-hand, why speededness is not to be dismissed by the law school admissions industry as a problem affecting sub-populations – those who can easily be ignored by adjusting a few definitions. We demonstrated that speededness impacting a small number of test-takers easily robs a standardized test of its power characteristics, so that it becomes a mixed power and speed instrument whose reliability and validity cannot be easily measured and, certainly, is not the test of ability that test-takers have a right to expect, pursuant to contract law, tort law and Civil Rights law. We outlined a three-count complaint against the stakeholders, including a law school that rejected an applicant, in part, due to a low LSAT score. The ABA, DOE, LSAC and US News could also be defendants. Finally, we showed that LSAC will have to prove that the LSAT and the surrounding test administration system meet
scientific standards of admissibility if LSAC is to have any chance of prevailing. We
demonstrated why defeat for the defendants is more likely.

This paper is the result of a law student’s research. Experienced lawyers will have to sift
through the above and determine whether they can drill down and/or extend the proofs and
strategy discussed here. Only through the diligent efforts of numerous attorneys and their test
clients can the LSAT be removed as the impediment to diversity in law schools that it so clearly
is today.

4.1 Caveat

In the eleventh hour of the final editing of this paper, the author uncovered a single line
from a single author suggesting that two sections of the LSAT are intentionally speeded.136
Since our assumptions are that the LSAT is a test of ability and not speed, intentional speeding
could jeopardize our findings for half of the test. Fortunately, this is not quite the case.

First, we captured evidence, above, that the LSAT was not intentionally speeded, per the
direct statements of LSAC and its researchers.137 Second, intentional speeding would have
unpredictable results and would challenge the validity of the LSAT. It seems improbable that
LSAT would risk its already low predictive validity values on such a venture. Thus, it is unlikely
that the LSAT is intentionally speeded.

Nonetheless, let us assume that two sections of the LSAT are intentionally speeded.
Then, the introduction of this new variable would actually work against LSAC and the LSAT, and
work to the advantage of those injured by the LSAT or who are otherwise opposed to it. LSAC

136 Stephen G. Sireci, Unlabeling the Disabled: A Perspective on Flagging Scores from Accommodated Test
137 LSAC, Cautionary Policies Concerning the LSAT, supra note 15, at 1; Schnipke & Scrams, supra note 15, at 2, 16.
would be required to demonstrate reliability and validity of its measure of the speed trait for which the sections allegedly test. This would entail statistical calculations for which there is no scientific consensus. It would also require evidence of testability, including successful trials, which LSAC cannot demonstrate because it cannot consistently measure speededness under current testing conditions. It would also have to explain how a peer review could have approved adding speeded components to a test already criticized for being a racially disparate speed drill.

Through discovery, the plaintiffs would learn why the change was implemented, when the intentional speeding was approved, who approved it, which sections are involved, how the change has been implemented, how much speededness was introduced, how often the sections are speeded, how the speed trait is measured, whether the outcomes have been racially or otherwise demographically differentiated, how the changes were publicized, etc. As well, depending on the degree of speededness introduced, much of the analysis in this study remains relevant, because Gulliksen’s work is based largely on approximations.

A final point is that there are likely errors in the above calculations, their underlying assumptions, and/or the inferences drawn from them. The concepts should still be valuable, however, and should be explored further, given the significance of the issues surrounding law school admissions. At worst, this paper would simply serve as an introduction to the relevant issues.
Appendix A: Harold Gulliksen’s Speededness Model

In this appendix, we examine Gulliksen’s model in more detail. We captured the conclusions elsewhere in this report.

Recall that Gulliksen’s model can be expressed using:

- $R$ – items correctly answered
- $W$ – items incorrectly answered or omitted after consideration
- $U$ – items unattempted (not considered, not reached within the time limit)
- $N$ – the total number of items
- $E$ – items NOT marked correctly (unattempted, incorrect, omitted after consideration)

$$E = W + U$$
$$R = N - E$$

Taking the standard deviation, $S$, of each side of the equation $E = W + U$, yields the following relationship between items NOT marked correctly and items unattempted:

$$S^2_{(W+U)} = S^2_W + S^2_U + 2r_{WU}S_W S_U$$

Gulliksen concluded that speededness of a test can be determined by comparing the standard deviations of unattempted items ($S_U$) and items NOT marked correctly ($S_E$ or $S_{(W+U)}$). When all test items are completed, $U = 0$, $S^2_U = 0$, $2r_{WU}S_W S_U = 0$, therefore, $S^2_{(W+U)} = S^2_W$. This means that the variance (or standard deviation) of items NOT marked correctly ($W$) is due ONLY to items

---

138 Rindler, supra note 3, at 263.
139 Id., at 263; Gulliksen, supra note 5, at 232. Note that the standard deviation is a measure of the scatter of values around an average. The standard deviation of a sum, $(x+y)$, can be expressed as $s^2_{(x+y)} = s^2_x + s^2_y + 2rs_x s_y$. See Gulliksen, supra note 5, at 425. The variable “$r$” is the correlation coefficient, which is a measure of how closely related two variables are. Its value is always >= -1 (perfect negative correlation) and <= +1 (perfect positive correlation). See Gulliksen, supra note 5, at 424.
140 Rindler, supra note 3, at 263.
141 Id., at 263. The variance is the square of the standard deviation and is often used to express relationships between standard deviations, for mathematical ease. The term “$r$” is the correlation coefficient, a measure of how well two variables track each other.
fully considered.\textsuperscript{142} This is the definition of a power test. Hence, when all test items are completed, the test can be considered a power test, which is completely unspeeded.\textsuperscript{143}

Similarly, when the test is wholly speeded, all questions are answered correctly, there is no contribution of items NOT marked correctly (W), hence $S^2_W = 0$ and $2r_{WU}S_WS_U = 0$.\textsuperscript{144} Of course, there will be items unattempted (U $<> 0$ and $S^2_U <> 0$). So, for wholly speeded tests, $S^2_{(W+U)} = S^2_U$.\textsuperscript{145}

Since the mixed speed and power test exemplified by the LSAT will approach a power test only when $S^2_U$ approaches zero, leaving $S^2_{(W+U)} = S^2_W$, and it will approach a speed test only when $S^2_W = 0$, leaving $S^2_{(W+U)} = S^2_U$, Gulliksen proposed examining the ratios $S_U/S_{(W+U)}$ and $S_W/S_{(W+U)}$ to gauge where along a spectrum of speededness a test is.\textsuperscript{146} It turns out that as $S_U/S_{(W+U)}$ approaches 0 (when U approaches 0\textsuperscript{147}), $S_W/S_{(W+U)}$ approaches 1, clearly indicating a power test, and vice versa.\textsuperscript{148} These fractions identify a speed test or a power test, then, only if one of them approaches 0, but not both.\textsuperscript{149}

The relevant questions are (a) at which points do these fractions indicate that the test being measured suffers no effects of time pressure (is a power test, i.e., is not speeded); and (b) whether the fractions indicate gradations of speededness such that there is a cutoff beyond which speededness has destroyed the validity and/or reliability of the test. Gulliksen made several key observations.

\textsuperscript{142} Id., at 263.
\textsuperscript{143} Id.
\textsuperscript{144} Id.
\textsuperscript{145} Id.
\textsuperscript{146} Id.
\textsuperscript{147} This would produce vanishingly small scatter, hence standard deviation approaching 0.
\textsuperscript{148} Rindler, supra note 3, at 263.
\textsuperscript{149} Id.
First, when \( W \) and \( U \) are perfectly negatively correlated \((r_{WU} = -1)\), the standard deviation of both \( W \) and \( U \) could be greater than that of \( E \) \((= W + U)\).\(^{150}\) This is because \( r_{WU} = -1 \) implies \( S^2(W+U) = S^2_W + S^2_U - 2S_W S_U \), which is an expression of the standard deviation of \( W - U \) or \( U - W \).\(^{151}\) That is \( r_{WU} = -1 \) implies \( S(W+U) = S_W - S_U \) or \( S(W+U) = S_U - S_W \).\(^{152}\) Hence \( S_U \) and \( S_W \) could be orders of magnitude greater or less than \( S(W+U) \), or one could be an order of magnitude greater than the other – even though they closely track each other \((r_{WU} = -1)\).\(^{153}\) Thus, neither \( S_U \) nor \( S_W \) can be used to predict the magnitude of the other, thereby requiring any analysis to evaluate both \( S_U \) and \( S_W \).\(^{154}\)

Second, only when \( W \) and \( U \) are perfectly correlated \((r_{WU} = -1 \text{ or } +1)\) AND one of these ratios, \( S_U/S(W+U) \) or \( S_W/S(W+U) \), is less than 0.1 could the equations certify that a test was either primarily a speed test or primarily a power test.\(^{155}\) From the above, where \( S_U/S(W+U) < .10 \), \( r_{WU} = -1 \), then it follows that \( S(W+U) = S_W - S_U \), \( |S(W+U)| = |S_W - S_U| \), \( S_W - S_U > S(W+U) > S_U - S_W \) and \( 1 = S_W/S(W+U) - S_U/S(W+U) \) and, finally, Gulliksen's speed and power equations become:\(^{156}\)

**Speed Test** if \( S_W/S(W+U) \) is small, then: \( 1 + S_W/S(W+U) > S_U/S(W+U) > 1 - S_W/S(W+U) \)

**Power Test** if \( S_U/S(W+U) \) is small, then: \( 1 + S_U/S(W+U) > S_W/S(W+U) > 1 - S_U/S(W+U) \)

\(^{150}\) Gulliksen, supra note 5, at 232.

\(^{151}\) See Id., at 425.

\(^{152}\) Id., at 232.

\(^{153}\) Id.

\(^{154}\) Id.

\(^{155}\) Id.

\(^{156}\) Id.
Gulliksen considered the 0.1 maximum value as acceptable because these equations narrow the band of possible values for the complementary ratios significantly, as in:  

\[ \text{Power Test if } S_U/S_{(W+U)} = 0.1, \text{ then: } 1.1 > S_W/S_{(W+U)} > 0.9 \]

\[ \text{Power Test if } S_U/S_{(W+U)} = 0.01, \text{ then: } 1.01 > S_W/S_{(W+U)} > 0.99 \]

Gulliksen stated:

If \( su/sx = 0.1 \), [the ratio \( sw/sx \)] must lie between 0.9 and 1.1. If \( su/sx = 0.01 \), the ratio cannot be less than 0.99 nor more than 1.01. In such a case we have a test that is primarily a power test, in the sense that the test variance would not be changed much if the subjects were allowed to finish the test. At one extreme possibility, if they were allowed to finish, they would all get all the unfinished items wrong, in which event the new \( sx \), would equal the old one. At the other extreme no one would get any of the items wrong, in which case the new \( sx \) would be equal to the present \( sw \), which, as we have seen above, must be within 10 per cent of \( sx \) if the ratio \( su/sx = 0.1 \).

Unfortunately, as these two fractions depart from zero, interpretations of speededness become nearly impossible.\(^{158}\) Gulliksen, for instance, described a case where he obtained \( S_U/S_{(W+U)} \) of 0.75, but for this value, he showed that \( S_W/S_{(W+U)} \) could be as small as 0.25 or as large as 1.75, per the correlation of \( U \) and \( W \).\(^{159}\) In effect, this model is capable of reasonably determining whether the test is a power test or speed test only when its two core, quantitative assumptions are valid: \( U \) and \( W \) are perfectly correlated and the above ratios are less than 0.1.

This model predicts largely the obvious, that a partially speeded test has the characteristics of a power test when \( U \) is nearly 0. But, we know that this can be a fiction because it predicts the presence of a power test should a significant portion of the test-takers fail to reach a similar number of questions, bringing \( S_U \) close to 0, while average \( U \) could be far greater than 0. Such a test, therefore, might be substantially speeded. It also assumes that no

\(^{157}\) Id.
\(^{158}\) Rindler, supra note 3, at 263.
\(^{159}\) Id.
questions have been skipped, which is not realistic. The theory might be sound, but the practical applications are limited.

Gulliksen also uses his model to define reliability estimates as “the correlation between two parallel forms with equal variances.” He notes that a test-retest scenario is required to estimate reliability of a mixed speed and power test, while a split-half correlation can be performed on a pure power test. Thus, being able to use these reliability estimates depends on prior knowledge of the level of speededness of the tests, requires test-takers to sit for a second test when mixed power and speed are observed, and requires uniform difficulty in questions when a split-half correlation is calculated for a pure power exam. These assumptions and requirements are clearly not practical for the LSAT, given that LSAC rarely, if ever, investigates level of speededness and the pencil-and-paper administrations are far too inefficient to allow for re-testing. Gulliksen’s reliability estimates are not helpful, therefore, but some of Gulliksen’s interim observations are valuable, as discussed above.

---

160 GULLIKSEN, supra note 5, at 233-34. A parallel form is typically a second test that the test-taker takes after the original test and is frequently untimed. This is not desirable and has potential validity problems due to the psychology of individuals who are forced to take a second test that they know is for experimental purposes; the exams overlap in content, improving the number of accurate responses for the second test; etc. See Lu & Sireci, supra note 5, at 32.
161 Id. at 233-35. A split-half reliability estimate calculates correlation coefficients using answers from only one test, but correlates half of the test with the other (as in odd vs. even).
162 Id. at 233-35.
Appendix B: Deriving Test Speededness from Data Measuring Individual Speededness

Recall that Gulliksen characterized the relationship between the number of questions (N), the number of questions unanswered (U), the number answered incorrectly (W), their respective standard deviations (SU and SW), and the correlation coefficients (r) as:

\[ S^2_{W+U} = S^2_W + S^2_U + 2r_{WU}SWSU \]

Gulliksen explicitly noted that both definitions for pure power and pure speed tests, as well as approximations for power and speed tests can be formulated by removing complementary contributions from this equation. He noted that a pure power test can be represented as:\(^{163}\)

\[ S^2_{W+U} = S^2_W \]

and a power test is approximated when SU approaches 0 as SW becomes increasingly non-zero.

A pure speed test can be represented as:\(^{164}\)

\[ S^2_{W+U} = S^2_U \]

and a speed test is approximated when SW approaches 0 as SU becomes increasingly non-zero.

The relationship between speededness for an individual test-taker and that of the entire test has, thus, already been expressed by Gulliksen by his demonstration of how a single test-taker contributes to the standard deviations in the above three equations. That contribution is on the order of \( \sqrt{\sum(U_i-M_U)^2/T} \), where each test-taker is identified by a different index i, T is the total number of test-takers, each test-taker’s number of unanswered questions is Ui, and the mean of U is denoted by MU.\(^{165}\)

---

163 U=0 for all test-takers, which means that \( S^2_U = 0 \), which means that \( 2r_{WU}SWSU = 0 \).
164 W=0 for all test-takers, which means that \( S^2_W = 0 \), which means that \( 2r_{WU}SWSU = 0 \).
165 GULLIKSEN, supra note 5, at 424. He provides the definition of standard deviation, SU, as \( S_U^2 = \sum(U_i-M_U)^2/T \). We use T and not N because N is already defined as the number of questions in the test. We ignore the T-1 correction, for mathematical convenience; given the population size of our illustrations (100,000), the correction is of little
We can now use these definitions to illustrate the relationship between individual speededness and test speededness.

6.1.1.1 Illustration #1: how test speededness is impacted by individual speededness – 80% cutoff, unspeeded test-takers have no wrong answers, speeded answer 25% wrong

Consider the scenario in which 100,000 people took an LSAT with 100 questions (N=100). 80% of test-takers (80,000) had no unanswered questions \( U_i = 0 \) and no incorrect answers \( W_i = 0 \). Each of the 20% who did not complete the exam only completed 75% of the exam (75 questions), failed to answer 25 questions \( U_i = 25 \), and answered 25 incorrectly \( W_i = 25 \). The mean of \( W \), therefore, is 5 and mean of \( U \) is 5.

As shown above, the standard deviation of \( U \) is:

\[
\sqrt{\frac{\sum (U_i - M_U)^2}{T}} = \sqrt{\frac{(80,000*(0-5)^2 + 20,000*(25-5)^2)}{100,000}}
\]

\[
= \sqrt{0.8*25 + 0.2*400}
\]

\[
= \sqrt{100} = 10
\]

166 This, actually, cannot be measured reliably outside of computer-based testing, due to rapid guessing behavior on a test that does not penalize for incorrect answers. From a power test perspective, such answers are not answers taken after fully considering the questions, so their corresponding questions count as unanswered.

167 This is also impossible to measure reliably for pencil-and-paper tests, as indicated above.

168 This would almost certainly be due to lack of time, since there is no penalty for guessing incorrectly on the LSAT.
Note that the 80% (unspeeded) portion of the test-takers contribute only 20% of the total standard deviation\(^{169}\) and the 20% (speeded) contribute approximately 80%. This is because the 80% effectively defined the mean and, therefore, their contributions are much closer to the mean. The difference of squares associated with them is likely to be much smaller than that associated with the 20%. Hence, the 20% are likely to have a disproportionate impact on the calculation of standard deviation. In this illustration, the 20%, for whom the test was speeded, dominate the calculation with 8 times their anticipated contribution.

The standard deviation of \(W\) is:

\[
\text{sqrt}(\sum (W_i - M_U)^2/T) = \text{sqrt}((80,000*(0-5)^2 + 20,000*(25-5)^2)/100,000) = 10
\]

Note that the 80% (unspeeded) portion of the test-takers contribute only 20% of the total standard deviation\(^{170}\) and the 20% (speeded) contribute approximately 80%.

The standard deviation of \((W+U)\), assuming \(r = 1\) (perfect correlation, reasonable considering fact pattern):

\[
S^2_{(W+U)} = S^2_W + S^2_U + 2r_{WU}S_WS_U
\]

\[
= 100 + 100 + 2*10*10
\]

\[
= 400
\]

\[
S_{(W+U)} = 20
\]

Gulliksen noted that a pure power test has \(S_U/S_{(W+U)} = \text{max of 0.1},\) and a pure speed test has \(S_W/S_{(W+U)} = \text{max of 0.1}.\) However, in this scenario, with \(r = 1,\) \(S_U/S_{(W+U)} = S_W/S_{(W+U)} = 10/20 = 0.5.\)

This illustrative test is clearly not an approximation to a power test.

---

\(^{169}\) That is \((0.8*25)/(0.8*25 + 0.2*400) = 20/(20 + 80) = 20\%\). This is not completely accurate because the contribution is immediately constrained by the square root operation, but the relative magnitudes are still helpful in seeing how the individual levels of speededness affect the overall calculations for test speededness.

\(^{170}\) That is \((0.8*25)/(0.8*25 + 0.2*400) = 20/(20 + 80) = 20\%\). This is not completely accurate because the contribution is immediately constrained by the square root operation, but the relative magnitudes are still helpful in seeing how the individual levels of speededness affect the overall calculations for test speededness.
6.1.1.2 Illustration #2: how test speededness is impacted by individual speededness – 90% cutoff, unspeeded test-takers have no wrong answers, speeded answer 25% wrong

Consider the scenario in which 100,000 people took an LSAT with 100 questions (N=100). 90% of test-takers (90,000) had no unanswered questions (Uᵢ = 0) and no incorrect answers (Wᵢ = 0). Each of the 10% who did not complete the exam only completed 75% of the exam (75 questions), failed to answer 25 questions (Uᵢ = 25), and answered 25 incorrectly (Wᵢ = 25). The mean of W, therefore, is 2.5 and mean of U is 2.5.

As shown above, the standard deviation of U is:

\[
\text{sqrt}(\Sigma(Uᵢ - \mu_U)^2/T) = \text{sqrt}((90,000* (0-2.5)^2 + 10,000*(25-2.5)^2)/100,000)
\]

\[
= \text{sqrt}(0.9*6.25 + 0.1*506.25)
\]

\[
= \text{sqrt}(5.625 + 50.625)
\]

\[
= \text{sqrt}(56.25) = 7.5
\]

Note that the 80% (unspeeded) portion of the test-takers contribute only 9% of the total standard deviation and the 10% (speeded) contribute approximately 91%.

The standard deviation of W is:

\[
\text{sqrt}(\Sigma(Wᵢ - \mu_W)^2/T) = \text{sqrt}((90,000* (0-2.5)^2 + 10,000*(25-2.5)^2)/100,000)
\]

\[
= 7.5
\]

---

171 This, actually, cannot be measured reliably outside of computer-based testing, due to rapid guessing behavior on a test that does not penalize for incorrect answers. From a power test perspective, such answers are not answers taken after fully considering the questions, so their corresponding questions count as unanswered.

172 This is also impossible to measure reliably for pencil-and-paper tests, as indicated above.

173 This would almost certainly be due to lack of time, since there is no penalty for guessing incorrectly on the LSAT.

174 That is (0.9*6.25)/(0.9*6.25 + 0.1*506.25) = 5.65/(5.625 + 50.625) = approx. 9%. This is not completely accurate because the contribution is immediately constrained by the square root operation, but the relative magnitudes are still helpful in seeing how the individual levels of speededness affect the overall calculations for test speededness.
Note that the 90% (unspeeded) portion of the test-takers contribute only 9% of the total standard deviation\textsuperscript{175} and the 10% (speeded) contribute approximately 91%.

The standard deviation of (W+U), assuming $r = 1$ (perfect correlation, reasonable considering fact pattern):

$$S^2_{(W+U)} = S^2_W + S^2_U + 2r_{WU}S_W S_U$$

$\Rightarrow 56.25 + 56.25 + 2*7.5*7.5$

$\Rightarrow 225$

$S_{(W+U)} \Rightarrow 15$

Gulliksen noted that a pure power test has $S_U/S_{(W+U)} = \max$ of 0.1, and a pure speed test has $S_W/S_{(W+U)} = \max$ of 0.1. However, in this scenario, with $r = 1$, $S_U/S_{(W+U)} = S_W/S_{(W+U)} = 7.5/15 = 0.5$.

This illustrative test is clearly not an approximation to a power test.

6.1.1.3 Illustration #3: how test speededness is impacted by individual speededness – 90% cutoff, unspeeded and speeded test-takers have approximately same number wrong

Consider the scenario in which 100,000 people took an LSAT with 100 questions (N=100). 90% of test-takers (90,000) had no unanswered questions\textsuperscript{176} ($U_i = 0$) and average of 10 incorrect answers ($W_i = 10$)\textsuperscript{177}. Each of the 10% who did not complete the exam only completed 85% of the exam (85 questions), failed to answer 15 questions ($U_i = 15$)\textsuperscript{178}, and answered 9 incorrectly ($W_i = 9$). The mean of W, therefore, is 9.9 and mean of U is 1.5.

As shown above, the standard deviation of U is:

$$\sqrt{(\sum (U_i - Mu)^2)/T} = \sqrt{(90,000*(0-1.5)^2 + 10,000*(15-1.5)^2)/100,000}$$

\textsuperscript{175} That is (0.9*6.25)/(0.9*6.25 + 0.1*506.25) = 5.65/(5.625 + 50.625) = approx. 9%. This is not completely accurate because the contribution is immediately constrained by the square root operation, but the relative magnitudes are still helpful in seeing how the individual levels of speededness affect the overall calculations for test speededness.

\textsuperscript{176} This, actually, cannot be measured reliably outside of computer-based testing, due to rapid guessing behavior on a test that does not penalize for incorrect answers. From a power test perspective, such answers are not answers taken after fully considering the questions, so their corresponding questions count as unanswered.

\textsuperscript{177} This is also impossible to measure reliably for pencil-and-paper tests, as indicated above.

\textsuperscript{178} This would almost certainly be due to lack of time, since there is no penalty for guessing incorrectly on the LSAT.
= \sqrt{0.9 \times 2.25 + 0.1 \times 182.25} \\
= \sqrt{2.025 + 18.225} \\
= \sqrt{20.25} \\
= 4.5

Note that the 90% (unspeeded) portion of the test-takers contribute only 10% of the total standard deviation\(^{179}\) and the 10% (speeded) contribute 90%. This indicates the intensity of the impact of speededness and helps to explain why speededness for a few individuals really does impact speededness for the test. It also explains why an arbitrary cutoff at 80% or 90% bears little relationship to overall or individual speededness.

The standard deviation of \(W\) is:

\[
\sqrt{\frac{\sum(W_i - M_U)^2}{T}} = \sqrt{\frac{(90,000 \times (10-9.9)^2 + 10,000 \times (9-9.9)^2)}{100,000}} \\
= \sqrt{\frac{0.9 \times 0.01 + 0.1 \times 0.01}{0.009 + 0.001}} \\
= \sqrt{0.009 + 0.001} \\
= \sqrt{0.01} \\
= 0.1
\]

Note that the 90% (unspeeded) portion of the test-takers contribute 90% of the total standard deviation\(^{180}\) and the 10% (speeded) contribute 10%. This perfectly proportionate contribution is to be expected because the number incorrect responses for members in each group is nearly identical and exactly 1 from either side of the mean. Thus, each group should contribute

\(^{179}\) That is \((0.9 \times 2.25)/(0.9 \times 2.25 + 0.1 \times 182.25) = 0.2025/(2.025 + 18.225) = 0.10\%.\) This is not completely accurate because the contribution is immediately constrained by the square root operation, but the relative magnitudes are still helpful in seeing how the individual levels of speededness affect the overall calculations for test speededness.

\(^{180}\) That is \((0.9 \times 0.01)/(0.9 \times 0.01 + 0.1 \times 0.01) = 0.009/(0.009 + 0.001) = 0.90\%.\) This is not completely accurate because the contribution is immediately constrained by the square root operation, but the relative magnitudes are still helpful in seeing how the individual levels of speededness affect the overall calculations for test speededness.
roughly equally to the distance-from-the-mean calculations which are at the heart of the standard deviation formula. Qualitatively, this indicates that the groups introduce the same amount of error or dispersion into the calculations.

The standard deviation of \((W+U)\), assuming \(r = 1\) (perfect correlation, reasonable considering fact pattern):

\[
S^2_{(W+U)} = S^2_W + S^2_U + 2r_{WU}S_W S_U
\]

\[
= 0.01 + 20.25 + 2*0.1*4.5
\]

\[
= 21.16
\]

\[S_{(W+U)} = 4.6\]

Gulliksen noted that a pure power test has \(S_U/S_{(W+U)} = \text{max of } 0.1\), and a pure speed test has \(S_W/S_{(W+U)} = \text{max of } 0.1\). However, in this scenario, with \(r = 1\), \(S_U/S_{(W+U)} = 4.5/4.6 = 0.978\) and \(S_W/S_{(W+U)} = 0.1/4.6 = \text{approx. } 0.022\). This illustrative test is clearly an approximation to a speed test and not a power test. Note that the ETS rule fails decisively to detect speededness here.

One explanation is that this scenario differentiates between speeded and unspeeded populations solely in their speed of response, but not in their subject matter ability. This might be the equivalent of imposing a penalty on incorrect answers, so that test-takers will only answer questions they have fully considered. That such a test approximates a speed test is disturbing because it means that similarly talented and capable test-takers will score dramatically differently due to the difference in number of questions they were able to consider fully: \((U_i = 0)\) vs. \((U_i = 15)\).

6.1.1.4 Conclusion

This exercise demonstrated that:

- Speededness of a test, as defined by the departure of the standard deviations of \(W\) and \(U\) from characteristics of a power test, can be described mathematically, assuming that
the quantities W and U are observed through computer-based or other IRT-supportive testing;

- A proportionate representation for speededness, such as that used by ETS, bears little resemblance to the models of speededness that define power and speed tests because they indicate nothing about standard deviation and correlation (we chose S and r in the above for illustrative reasons, but nothing in the ETS rule constrained those choices);

- In an illustration of an LSAT exam in which 90% of the test-takers experienced no speededness, while 10% did not complete 15% of the test due to time constraints, we held ability roughly equal for the portions of the test both groups completed (W were approximately equal). In this case, the test appeared to have the characteristics of a speed test and bore little resemblance to those of a power test. Fundamentally, this suggests that students of equal ability would be scored differentially, exclusively due to speededness. One of the LSAC models suggested “slight” speededness, pursuant to Wightman & Muller, while the ETS model suggested no speededness. Both models were incorrect;

- In an illustration of an LSAT test that is a pure power test for 80% of the test-takers and maximally speeded, per the ETS standard, for the remaining 20% (they did not reach 25% of the test), the test has no characteristics that approach that of a power test. This actually means that the ETS standard predicted speededness correctly in this instance. However, it also means that the level of speededness was so great that the test could no longer be considered a test of ability. This suggests that, in some circumstances, the ETS
standard might not mark the beginning of speededness in a harmless, academic
characterization, but the pronounced effects of speededness;

- The contribution to the standard deviations of W, U and (W+U) from the population for
  which a test is speeded can far outweigh that of the contributions of the non-speeded
test-takers, when there is sufficient disparity in the impacts of speededness;
- The impacts of individual test-takers, who found the test speeded, on the speededness
  of the entire test is non-linear, but it can potentially outweigh the combined
  contributions of those for whom the test was not speeded by an order of magnitude.

Essentially, a test is speeded only if it is speeded for individual test-takers. Speededness
can be measured by standard deviations, which are non-linear entities. But, the ETS-like rules
for speededness are governed by linear percentages and proportions that in no manner
approximate standard deviations. The ETS-like rules cannot determine test speededness.