FINAL REPORT

An Independent Review of Educational Testing Service's Study on the Predictive Validity of the Graduate Record Exams for Making Law School Admissions Decisions

Won-Chan Lee and Robert L. Brennan

The Center for Advanced Studies in Measurement and Assessment

The University of Iowa

September 7, 2021

Table of Contents

Executive Summary

The Center for Advanced Studies in Measurement and Assessment (CASMA) finds the results in the ETS Research Report ETS –RR-18-26 (Klieger et al., 2018) to be an insufficient basis for a clear recommendation that the GRE and LSAT can be used interchangeably and successfully for admissions to any/all law schools.

If certain plausible assumptions are made, however, we think it is possible that the GRE and LSAT might be used defensibly and interchangeably for law school decisions. This leads us to recommending that a Pilot Study be undertaken with a sample of law schools. The Pilot Study would need to be conducted under conditions that mirror, as closely as possible, the circumstances that would prevail if the ABA ultimately endorses an "either/or" policy with respect to using both the GRE and the LSAT. The Addendum to this report provides more detailed suggestions for conducting such a Pilot Study.

If the above recommendation is not accepted, and the ABA adopts an either/or policy for the GRE and LSAT, then we strongly recommend that the decision be revisited after 3-5 years, at which time adequate data should be available to assess the extent to which the either/or policy has been successful.

Background

The Center for Advanced Studies in Measurement and Assessment (CASMA) was hired by the American Bar Association (ABA) to independently review a study carried out by the Educational Testing Service (ETS) that looked at the predictive validity of the Graduate Record Examination (GRE) for making law school admission decisions. ETS published a study, entitled "*The Validity of GRE® General Test Scores for Predicting Academic Performance at U.S. Law*

Schools" (Klieger et al., 2018), as part of their on-going research report series. This report is referred to as the "ETS Report" or the "Report" (note uppercase R) throughout this document.

This review is a report to the ABA as the primary audience and attempts to address the ABA's principal question of whether the GRE can be used interchangeably with the LSAT for admissions to any/all law schools. In other words, *our evaluation is focused on the possible "either/or" policy that the ABA may consider with respect to using either GRE or LSAT for making law school admissions decisions*.

Parts of our review of the ETS Report employ basic ideas in Kane's (2006) validation framework. In particular, we view validation as a continual process of making claims and gathering evidence to support those claims. Under this framework, the type and amount of evidence required to support a given claim depend on the claim itself. Stronger evidence may be required to support the most impactful claims while weaker evidence may suffice for less consequential claims.

Note that it was not our contractual responsibility to examine LSAT or other ETS documents that might potentially be used to further support use of the GRE in the context of law school admissions. One exception was a confidential memo entitled "*GRE–LSAT Concordance*" (dated July 5, 2021) that we received from ETS while we were working on this version of this report. Elsewhere in this report we refer to the confidential memo as the "ETS Memo."

An Addendum to this report was written by the second author of this report. The Addendum provides detailed discussions of various technical issues concerning the contents of this report---particularly issues concerning concordance of GRE and LSAT scores, as well as other important matters regarding use of the GRE under an "either/or" policy.

We begin this report by summarizing the ETS Report. Then, we offer a set of important claims and evidence provided directly or indirectly in the ETS Report along with some suggestions to further strengthen the validity argument for the use of the GRE as a suitable law school admissions test. Finally, we discuss additional considerations associated with accepting the GRE as a valid law school admissions exam and make recommendations for next steps.

Review of the ETS Study

One perspective on the study in the ETS Report is that it was designed to answer three main questions:

(1) What is the predictive validity of each GRE section?

(2) What is the predictive validity when different section scores are combined?

(3) What is the predictive validity when GRE section scores are considered in addition to undergraduate grade point average (UGPA)?

The GRE used in the study was revised in 2011 and was delivered as a computerized multi-stage adaptive test (MST) for its verbal and quantitative sections and as a linear computer-based test (CBT) for its analytical writing section.

Since the GRE is being considered for use interchangeably with the LSAT, ETS studied the GRE's predictive performance compared to that for the LSAT. The LSAT and the GRE differ with respect to their administration modes. In particular, the LSAT is a linear CBT, whereas the GRE is mainly composed of adaptive sections. Also, the LSAT and GRE differ with respect to the number of reported scale scores. Typically, an applicant who takes the GRE is provided with *three* separate scale scores: verbal (V), quantitative (Q), and analytical writing (AW). By contrast, LSAT reports a *single* scale score for each examinee. Presumably, to

accommodate the fact that a GRE composite score is not typically reported, ETS considered certain composite GRE scores for use in their study. For example, one composite reflected an equally weighted sum of just the GRE-V and GRE-Q; another composite used different weights for GRE-V and GRE-Q.

The ETS study was based on a sample of 1,587 students spanning 21 U.S. law schools accredited by the ABA. A number of these schools were self-selected to participate in the study (exact numbers were not provided). Of these 21 schools, 15 were private and 6 were public. Ten of the 21 schools had students who took the GRE prior to matriculation (these tended to be larger schools) and the remainder had students who took the GRE after matriculation, on a good-faith effort accompanied by monetary compensation. Schools were sampled based on a stratification of five geographic regions (northeast, south, southwest, midwest, and west) and three school selectivity levels (using median LSAT and UGPA of entering students).

Several analyses were carried out to examine the predictive validity of the GRE—many of which were correlational in nature. The criterion used throughout the study was first year law school GPA (LGPA). Since roughly half of the sample took the GRE prior to entering law school and about half took it after enrollment into law school, ETS first examined the similarities among these two groups to determine if they could be combined sensibly for subsequent analyses. To do this, ETS computed two sets of validity coefficients (i.e., correlations between the LSAT and LGPA and between the GRE and LGPA) for each group and examined the overlap in their confidence intervals. ETS concluded that the validity coefficients were similar across groups and, therefore, concluded that the two groups could be combined in subsequent analyses.

The bulk of the analyses in the ETS study were correlational and meta-analytic in nature. First, internal consistency reliability of the GRE was compared against that for the LSAT and

was found to be similar. For GRE composite scores, the internal consistency reliability was reported as .96 and was deemed comparable to LSAT reliability estimates that range between .90 and .95.

Validity evidence was then gathered using analyses such as correlations, ordinary least squares regression, meta-analysis techniques, contingency table analyses that resembled classification consistency, and hierarchical linear regression. Correlations between LGPA and composite GRE scores were similar to the correlations between LGPA and LSAT scores. Meta-analyses showed that the estimated mean true score correlation between the GRE and LGPA (i.e., .53 and .54 for GRE V+Q+AW and GRE V+Q, respectively) was similar to that between the LSAT and LGPA (i.e., .55). Contingency table analyses showed that among students who scored in the top third of the GRE, roughly 45% also fell in the top third of their law school class. This implies that roughly 55% of students fell below the top third of their law school class. Similar contingency table results were not reported for the LSAT. Lastly, ETS carried out hierarchical linear modeling to determine if validity coefficients differed across school selectivity levels. They found a statistically significant interaction for school selectivity but its corresponding effect size was small and not practically significant. In turn, it was concluded that validity coefficients were similar across schools contained in the study's sample and across school selectivity levels. In general, the analyses contained in the ETS Report provide some evidence for the predictive validity of the GRE as a predictor of first-year law school performance.

Validity Arguments

While there are numerous claims that could be made in a predictive validity study for the GRE, there are a handful that CASMA finds to be most critical, including (1) characteristics of the GRE, (2) predictive power, (3) generalizability of results, and (4) concordance between the GRE and LSAT scores. Each of these categories is discussed next. (Additional details, especially for 4, are discussed in the Addendum.)

Characteristics of the GRE

Although the ETS Report claims that the content of the GRE relates to pre-requisite knowledge required to be successful in law school, content validity is not given much consideration. Rather, the primary focus of the ETS Report is on predictive validity. While content might be only indirectly related to predictive validity, content validity is a standalone component that is very important to the overarching validation process.

In particular, CASMA argues that, in order to make an informed "either/or" decision, the ABA needs to know how the content of the GRE compares that of the LSAT. Obviously, content plays a critical role in all testing situations and certainly warrants attention in the current circumstances. An obvious way to evaluate the appropriateness of the GRE content would be to have law school subject matter experts review and compare the content specifications (typically operationalized by test blueprints) for both the GRE and the LSAT. One group of such experts might be those who advise the LSAC on the content of the LSAT. The ABA could, as well, choose other groups of content experts. Such a review (or reviews) would likely also consider the appropriateness of differences in content tested by both programs. Content considerations are discussed further in the Addendum.

Another topic often discussed in the psychometric literature but left out in the ETS Report is the differences in administration modes between the LSAT and GRE. At the time of the study, the GRE was a computerized test administered in a multi-stage format. By contrast, the LSAT was a paper test administered in a linear format. If the ETS study were conducted with the current version of the LSAT, which is now delivered as a computer-based test, conclusions might be slightly different. This last point naturally leads to the observation that validation is an on-going process of continuously making interpretive arguments with respect to score use and continuously collecting evidence to support such use. Accordingly, since LSAT and GRE mode differences are now less pronounced, it would seem advisable that another validity study be considered that examines the impact of mode differences.

Predictive Power

The fundamental argument made throughout the ETS Report is that performance on the GRE is indicative of future performance in law school. The evidence to support this claim is provided by correlational and meta-analytic analyses and in the similarity between validity coefficients for the GRE and LSAT with respect to predicting first year LGPA. The methodological approaches and magnitudes of coefficients appear to be in line with typical predictive validity studies. Both the GRE and the LSAT showed a similar level of predictive validity when used alone or combined with UGPA.

Often, when there is a need to determine whether a measurement procedure has predictive power, data are collected at two different timepoints. Ideally, the measurement procedure in question is given at a time point that precedes the time in which outcome of interest (i.e., first year LGPA in the case of the ETS study) is measured. Ideally, the measurement procedure and the criterion outcome are separated by some amount of time such as days, months,

or years, depending on the situation. In the case of the ETS study, for roughly half the sample, there was very little to no separation between the time the GRE was given and the time the criterion outcome (i.e., first year LGPA) was collected. Specific details regarding this timeline for those who took the GRE on a good faith effort were not provided in the study. For example, the content learned during the first year of law school might have affected GRE test performance for some participants. The impact of this on study results is unknown.

ABA standard 503-1 states that "A law school that uses an admission test other than the Law School Admission Test sponsored by the Law School Admission Council shall demonstrate that such other test is a valid and reliable test to assist the school in assessing an applicant's capability to satisfactorily *complete the school's program of legal education"* (emphasis added)." It is important to note that the above quote emphasizes collecting evidence for "*a*" law school, whereas the ETS Report provides results based on the combined data *over* law schools. This is not unusual since institution-specific data and results typically are treated as confidential and are not reported publicly. Accordingly, if the ABA adopts an "either/or" policy, CASMA recommends that the ABA encourage each individual law school to examine school-specific data.

Generalizability

For the ETS study to have utility for an ABA endorsement of an "either/or" policy, there should be evidence provided in the Report demonstrating that its findings are generalizable to the target population of *both* law schools and students. To address this, ETS used a stratified sampling design categorized by geographic location, school selectivity, school size, and public versus private school status. This resulted in a sample of 1,587 students from 21 of the 205 ABA accredited law schools. Essentially ETS claims this sample is representative of the target

population. Although a larger sample of schools might be desirable, ETS has some evidence to support a claim that the results of their study generalize to the population of law schools.

However, the representativeness of students is rather questionable in the following two senses. First, only matriculated students are represented in the data. That is, there are no data for students who applied to law school but were *no*t admitted. Second, only about half of the 1,587 students had preadmission GRE scores. GRE scores for the other students were collected in an ETS special study *after* the students were admitted. ETS deserves credit for collecting the data, but unfortunately the timing of the data collection (after admission) is inconsistent with the use of the GRE in an "either/or" environment which would involve taking the GRE or the LSAT *prior* to an admissions decision.

CASMA believes that the target-population limitations noted in the previous paragraph present a serious challenge to the generalizability of results for an 'either/or" ABA endorsement. ETS argues that similarities in validity coefficients for the pre-admission and post-admission GRE groups justify combining the two groups and claims that the combination represents the same target population. CASMA does not believe that these coefficients are sufficient evidence.

Also, ETS assumes that student demographic groups defined by race, ethnicity, and gender, for example, are also represented adequately and are comparable to what would be found in the target population. However, this assumption regarding student demographics was not investigated. It is difficult to support claims about the generalizability of the ETS study results in the absence of considering generalizability for at least some critical student subgroups.

Another aspect of the sampling design that affects generalizability is the unknown number of schools that participated in the study by means of self-selection. The ETS Report makes it clear that some schools asked to join the study after learning about the involvement of

the University of Arizona. It is unknown how many other schools self-selected and what level of systematic error this could have introduced.

Concordance

There needs to be a concordance relationship established between the LSAT and GRE scores if scores from both tests are to be used interchangeably by any institution under an "either/or" policy. As discussed extensively in the Addendum, concordance is substantially different from prediction. A concordance claim implies that it should be a matter of indifference to applicants, which test they take (i.e., either the LSAT or GRE) because the concorded scores will allow for the same interpretation and use. Put another way, applicants should have the same likelihood of being accepted to law school regardless of whether they take the LSAT or GRE, provided a concordance relationship is established between the LSAT and GRE scores. To accomplish this, typically an equipercentile concordance relationship is established such that particular GRE and LSAT scores are considered equivalent when they have the same percentile rank in the same target population. Typical correlational statistics alone are inadequate for this purpose.

In the ETS Memo, "*GRE–LSAT concordance*," ETS describes a procedure for relating the GRE and LSAT scores. A brief description of the ETS procedure, as well as an associated "tool" for performing computations, is available from: https://www.ets.org/gre/institutions/admissions/interpretation_resources/law_comparison_tool/. The above published description does not refer to the tool (or the resulting computations) as a concordance. By contrast, the confidential ETS Memo does refer to the procedure as a concordance. Here, we simply call this the ETS Procedure. It was established based on exactly the same data used for in the ETS Report. The ETS' GRE-to-LSAT Procedure may be useful for

some purposes; however, there are at least three issues that limit the utility of the ETS Procedure under an "either/or" policy: (a) data, (b) content, and (c) statistical/psychometric matters. Next, we briefly discuss these limitations.

First, the data used by ETS to develop the GRE-LSAT conversion were obtained from matriculated students only, not applicants. As a result, it is unknown if the conversion is applicable to non-admitted students. Second, as defined in the literature (e.g., Kolen & Brennan, 2014, ch. 10), concordance applies to scores on tests that are intended to measure *similar content* (or constructs) under similar measurement conditions. It is a composite of both GRE-V and GRE-Q that is linked to LSAT scores in the ETS Procedure. Clearly, the content of the ETS-proposed GRE composite and LSAT may not be comparable due, for example, to the existence of the quantitative section scores in the GRE composite. Note that we are not arguing for/against the adequacy of the content for GRE per se; rather, the focus here is on the adequacy of the statistical relationship between scores for the ETS-proposed GRE composite and scores for the LSAT. One consequence of using less-similar content is the potential differences in relationships across subgroups. Third, the statistical procedure used for the ETS Procedure involves use of a bivariate linear prediction concatenated with an equipercentile linking, which makes it quite different from conventional concordances as discussed in the literature. See the Addendum for a substantially more detailed discussion of the three matters discussed in this paragraph.

Final Considerations and Recommendations

CASMA was hired by the ABA to independently review a study carried out by ETS that looked at the predictive validity of the GRE for making law school admission decisions. It should be noted that this task did not include a review of the validity evidence pertaining to the

use of the LSAT for making law school admissions decisions. As a result, CASMA cannot make comparative statements regarding the level of validity evidence provided in support of the LSAT compared to that provided in support of the GRE. Nonetheless, all exams used for making college entrance decisions should be held to the same standards with respect to evaluating whether there is ample evidence provided to support intended uses and interpretations.

As previously mentioned, the ETS study is largely a predictive validity study that mainly features correlational and regression-based analyses. This leads us to the recommendation that a pilot study be undertaken with a sample of law schools to examine if the GRE and LSAT can be used successfully for admissions under typical conditions in an "either/or" environment. To be informative, such a study would need to be conducted under conditions that mirror, as closely as possible, the circumstances that would prevail if the ABA ultimately endorses an "either/or" policy with respect to using both the GRE and the LSAT. The Addendum to this report provides a more detailed set of recommendations for conducting such a pilot study.

If the above recommendation is not accepted, and the ABA adopts an either/or policy for the GRE and LSAT, then we strongly recommend that the decision be revisited after 3-5 years, at which time adequate data should be available to assess the extent to which the either/or policy has been successful.

As a final cautionary note, we note that it is customary for large-scale testing programs to undergo revisions based on the changing needs of the clientele they serve. CASMA sees no reason to believe that legal programs are an exception. Therefore, it would be prudent to consider whether the GRE can be flexible and reactive to the changing landscape of the legal profession and the needs of law school admissions committees. Such flexibility may be quite difficult to

attain for the GRE since it is already used for many different kinds of graduate programs that have substantially different needs and clienteles.

References

Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp 17-64). Westport, CT: American Council on Education/Praeger.

Klieger, D. M., Bridgeman, B., Tannenbaum, R. J., Cline, F. A., & Olivera-Aguilar, M. (2018). *The validity of* GRE® *General Test scores for predicting academic performance at U.S. law schools* (Research Report No. RR-18-26). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12213

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd ed.).* New York: Springer-Verlag.

Addendum:

A Discussion of Technical Issues

Relative to Using the GRE under a Possible ABA "Either/or" Policy for

Admission to Law Schools

Robert L. Brennan

September 7, 2021

This Addendum is part of CASMA's evaluation of the ETS Report entitled *"The Validity of GRE General Test Scores for Predicting Academic Performance at U.S. Law Schools"* (ETS RR--18-26).[1] Henceforth, this report is often referred to simply as the "Report" or the "ETS Report" (note the uppercase "R"). Note that the focus of CASMA's evaluation of the Report is restricted to the possible ABA "either/or" policy decision concerning the use of either GRE or LSAT for admission to law schools. This addendum should not be construed as an evaluation of the GRE per se, its use in other contexts, or the use of the ETS Report in other contexts. Since CASMA's evaluation is a report to the ABA, clearly the principal audience is the ABA, but the Addendum may also be informative or helpful to other stakeholders.

Here and elsewhere in this Addendum, I often use the "first-person point of view" (I, my, etc.) to refer to my personal opinion or perspective based on my knowledge and experience. Also, note that substantial parts of this Addendum are very specifically intended to be instructional, since I believe that many of the matters addressed here are not generally well understood, but some understanding of them is necessary to make an informed decision about an "either/or" policy. The overall goal is that the ABA in particular, as well as other stakeholders, have an enhanced understanding of the matters discussed here.

It is unknown to me *exactly* what ABA "approval" might mean. I assume here that the central focus is an "either/or" use in which law schools would be allowed (or required?) to accept either LSAT scores or GRE scores for consideration in admissions decisions. I refer to this generically as the "either/or" policy. Note that the word "policy" is purposeful. No amount of statistical or psychometric analyses can fully justify any "either/or" policy for all law schools, since there are differences among law schools, all test scores are fallible to some degree, and no two different testing programs (no matter how excellent they may be) test identical content.

Also, an "either/or" policy as stated in the above paragraph does not explicitly consider the use that might be made of GRE-V and/or GRE-Q. For example, the ETS Report argues that GRE-V and GRE-Q should be used jointly in a certain sense. This addendum argues that such joint use is likely to be problematic. On the other hand, it seems likely to me that there might be a defensible role for GRE-V. It also seems possible to me that GRE-Q might be used as an *auxiliary* source of

[1] Note that when this Addendum was being written, CASMA received a confidential memo from ETS entitled "GRE-LSAT Concordance" and dated July 5, 2021. My review of this memo led me to include an expanded discussion of some topics in this Addendum.

information, in a manner specified by the ABA or determined by each law school, although this Addendum does not address this matter in any detail. Note that GRE-AW is barely mentioned in this addendum, although it too might be used as *auxiliary* information in a manner to be determined by each law school. I take no position about what role, if any, GRE-AW might play in an "either/or" policy decision, or what role the LSAT writing sample might play.

This Addendum is definitely *not an overall evaluation* of the quality, value, or "validity" of the GRE or the LSAT testing programs. That is, the focus here is *not* on the individual programs per se, but rather the defensibility of their joint use in one specific context, namely, using the ETS proposed GRE composite (i.e., a weighed combination of GRE-V and GRE-Q) *interchangeably* with LSAT scores under an "either/or" law school admissions policy. Basically, there are two issues that need to be addressed to evaluate such use.

1. First, is the content of the tests comprising the ETS GRE composite "similar" to that of LSAT? Content matter experts need to consider this matter. I suspect that one important consideration will be the extent to which LSAT test content is comparable to a composite involving GRE-Q.
2. Even if the content were judged to be similar, adequate data still need to be collected and analyzed to support such use. The crux of the matter is the psychometric credibility (in the sense discussed later in this Addendum) of a "concordance" table that relates scores on the ETS GRE composite and scores on LSAT.

This Addendum addresses the following topics:

- Context and Use: An Overview
- Reliability and Validity Matters
    - Some Comments about Correlation Coefficients
    - Standard Errors of Measurement
- Prediction, Linking, Equating, Concordance and Related Matters
    - Fundamental Terminological Distinctions
    - The ETS GRE-LSAT Procedure
    - Concordance, Content Similarity, and Population Invariance: Basic Ideas
    - ACT-SAT Concordances: An Old Example
    - Some Implications for the "Either/or" ABA Policy Decision
    - Content and Related Considerations Revisited
- Recommendations
- Concluding Comments

As will become evident, one focus of this Addendum is "errors" in the psychometric and/or statistical sense, which is emphatically *not* the same notion as "mistakes" in the colloquial

sense.[2] Also, there is some repetition across sections to facilitate understanding of many interrelated but sometimes quite complicated matters.

Context and Use: An Overview

The LSAT has scale scores with a range of 120-180 (i.e., 61 scale-score points) with raw scores ranging from 0 to 101. Importantly, there are no generally reported subtest or section scores for LSAT. The GRE has three sections:

- GRE-V (a "verbal" test) with a scale-score range of 130 to170 (41 points) based on raw scores of 0 to 40;
- GRE-Q (a math test) with a scale-score range of 130 to 170 (41 points) based on raw scores of 0 to 40, approximately; and
- GRE-AW (a writing test with two prompts) with a scale-score range of 0 to 6 in units of .5 (13 points).

I do not have detailed knowledge about the admissions process for any law school. My understanding is that LSAT is accepted by all law schools, with some schools accepting either the GRE or LSAT. Presumably, the "either/or" schools must have some process that incorporates GRE scores into their admission decisions. I do not know if all "either/or" schools use the same process, or different processes. Even if they use the same process, I don't know what the process is.

For discussion purposes here, I make the simple assumption that each law school already has its own linear prediction of law school first-year grade point average (LGPA) based on LSAT scale scores and probably undergraduate grade point average (UGPA). To further simplify discussion, I overlook the role of UGPA. Doing so emphatically does *not* mean that UGPA is unimportant. Rather, I am trying to focus attention on the variables that appear to me to be are at the heart of an "either/or" policy decision, namely LSAT and GRE.

In the traditional linear-prediction statistical context, prediction "errors" reside in LGPA, not LSAT scores. That is, for any specific law school and any specific LSAT score, say 160, there likely are students with multiple different LGPAs. Prediction is a well-known and frequently practiced statistical methodology. In the testing context, however, there is an additional complexity not typically acknowledged in introductory statistics. Specifically, the predictor variable, observed LSAT scale scores in this discussion, contains errors of measurement. For example, if a student were to test multiple times with different forms of LSAT administered under similar circumstances, that person likely would obtain somewhat different LSAT scores. Such differences reflect "errors or measurement" for that examinee (discussed more fully later.)

[2] For an individual examinee, errors in the psychometric sense are the differences between observed scores (e.g., number of items correct on a single form of a test) and true score (the expected value of the examinee's observed scores over replications of the measurement procedure.) See Haertel (2006) for more details.

It follows that any linear prediction of LGPA from LSAT is doubly uncertain due to both *prediction errors* and *measurement errors* in the predictor variable, LSAT.

An important focus of this Addendum is the ETS proposed GRE-LSAT Procedure, which involves a weighted combination of *both* GRE-V and GRE-Q, with the weights being approximately .6 for GRE-V and .4 for GRE-Q. This procedure is described in more detail later on pp. 26 and 27. The procedure is sometimes abbreviated as the "ETS GRE-LSAT Procedure," the "ETS proposed Procedure," or simply the "ETS Procedure."

Because this procedure is complicated, however, it is very challenging to use it to explain a number of important matters. Therefore, I often refer to GRE-V, only, in discussing various psychometric and practical issues. The ETS proposed GRE-LSAT Procedure is explicitly identified when it is necessary to do so. Reference to GRE generally means the entire testing program.

In principle, there are at least two ways that GRE-V could be incorporated into the above prediction system: (a) use a two-variable prediction that would require all applicants to take *both* GRE-V and LSAT; or (b) let applicants take *either* GRE-V *or* LSAT.[3] The first alternative is unlikely for numerous practical reasons. The "either/or" alternative, which is the focus of this Addendum, may appear appealing, but it introduces complexities and additional potential sources of error.

Implementing the "either/or" alternative would require a three-step process:

1. Using data from an as-large-as-possible representative sample of law schools, identify applicants who took LSAT at one time and GRE at another not-too-different time.
2. Using the data from Step 1, create a concordance table (much more about this later) that provides an estimated one-to-one bi-directional relationship between LSAT and GRE-V scores.
3. Then, each law school could do the following.
    a) For each applicant who took LSAT, use that applicant's LSAT score in the extant LSAT-LGPA prediction equation for that school.
    b) For each applicant who took GRE-V, use the *concordance table* to get a concorded LSAT score for that applicant and plug it into the extant LSAT-LGPA prediction equation for the school.

Steps 1 and 2, above, introduce additional errors over and beyond those introduced in obtaining the LSAT-LGPA prediction equations for each school. Specifically, Step 1 is subject to *sampling error* in the sense that those who took both the GRE and LSAT may be a relatively small number of applicants. Step 1 is also subject to self-selection error in the sense that that those who chose to take (or were somehow encouraged to take) both tests might have different characteristics from those who chose to take only one test. These types of errors are associated, in part, with concordance being *potentially* population dependent. This implies, among other things, that the

[3] In practice, of course, applicants would almost certainly take all three sections of the GRE. Here, for instructional purposes, the discussion focuses on GRE-V, only, unless stated otherwise.

concordance might be different if it were based, say, on randomly selected examinees as opposed to self-selected examines, and/or the concordance might be different if it were based on only males, only females, only white applicants, only black applicants, etc. (Much more about this later.)

Sampling error is reduced when larger numbers of examinees take both tests. Self-selection error and population dependence, however, are not likely to be reduced by an increase in sample size. Almost certainly, for example, population dependence is largely the result of differences in *content* tested and/or administration procedures between the GRE-V and LSAT---hence, the importance of content considerations in concordance matters.

In the simplest case, GRE-V scale scores might be concorded with LSAT scale scores. If so, obtaining and using this concordance with any LSAT extant prediction equation would involve at least the following types of errors (directly or indirectly):[4]

1. Prediction error in the LSAT-LGPA prediction equation (i.e., variability in predicted LGPA for a given LSAT score);
2. Measurement errors in LSAT scores;
3. Measurement errors in GRE-V scores;
4. Sampling error involved in obtaining the data for a concordance;
5. Possible selection error (bias) in the data used to obtain the concordance; and
6. Errors in the sense of population-dependent differences in the content tested by LSAT and GRE-V---the importance of which is considered in much more detail later.

As noted previously, prediction error is typically quantified by a "standard error of estimate." Errors of types 2 and 3 are usually quantified as standard errors of measurement (SEMs) which are often associated with the term "reliability." Sometimes the word "validity" refers to the correlation involved in performing the prediction in 1, above. If so, the phrase "validity corrected for attenuation" (due to less than perfect reliability) likely would involve errors of types 1 and 2, 1 and 3, or 1-3 in a rather complicated expression. Errors of types 4-6 are *not* typically associated with "validity" as quantified by a correlation coefficient.[5]

In short, as discussed more fully next, the terms "reliability" and "validity," as well as actual estimates of them, can be very confusing, misleading, or largely uninformative for certain important issues discussed here. There is nothing wrong with reporting coefficients, but they are not substitutes for acknowledging, in some manner, different types of errors such as those in 1-6, above.

[4] Not all of these types of potential errors are given thorough attention here. The principal message is that there are multiple sources of error that need to be acknowledged, estimated if possible, and reduced to the extent possible under reasonable constraints.

[5] In the ETS Report, the only explicit reference to standard errors is with respect to the hierarchical linear model results in the Report's Appendix C.

Reliability and Validity Matters

The terms "reliability" and "validity" are generic characteristics typically associated with test scores. In the ETS Report, however, and in numerous other contexts, these terms are used to denote correlational-type statistics that have values that can range from 0 to 1. Such statistics are often functions of other more direct (and often more relevant) statistics, as noted in the previous section. Thus, estimates of "reliability" and "validity" often hide the magnitude of statistics directly relevant to the errors listed above. It is entirely possible for a correlationally-based estimate of reliability or validity to be (or seem to be) quite high when one or more types of error variance would be judged unacceptably large for some particular use.

Contrary to common belief (or wishes), there are no "magic" values for the magnitude of a reliability or validity correlation coefficient that certify interpretations such as "moderate," "acceptable," or "high." Even the *Standards for Educational and Psychological Testing* (2014) is, for the most part, notably silent about such matters.

Reliability in the generic sense refers to the *consistency* of scores over replications of a measurement procedure. So, for example, if examinees took the GRE-V twice and got similar scores, we would say that the scores were "reliable," at least to some degree. Estimating reliability requires making assumptions of one kind or another. Validity typically concerns the extent to which *claims* about scores are supported by *evidence*. (I prefer to call this "validation.")

Importantly, there is no such thing as "the" reliability or "the" validity of a *test* per se. It is *scores*, not tests that possess *different degrees* of reliability for different populations and/or under different assumptions or circumstances. Similarly, scores possess *different degrees* of validity, in the sense of different types and/or amounts of information, for *different interpretations and/or uses of test scores*, as well as for different populations.[6]

There are *numerous* procedures for *estimating* reliability and validity of test scores. Generally, these different procedures give *different* results. Often, the assumptions involved in obtaining estimates really do make a difference. So, any reference to "the" reliability or "the" validity can be doubly misleading and, in my opinion, often uninformative.

I acknowledge that vague use of terminology is common practice (even accepted practice) in many contexts but, except under highly restricted circumstances, doing so can be quite problematic. Of course, if readers are fully aware of the points raised here, abbreviated use of terminology is probably fine, but that is often not the case, in my experience.

For the most part, the ETS Report focuses on correlation coefficients when reference is made to "validity" or "reliability." For GRE-V and GRE-Q scores, I could not always determine whether coefficients were based on raw-scores (e.g., numbers of items correct) or scale scores (130 to 170).[7] I believe the GRE-V and GRE-Q scale scores are (or can be viewed as) non-linear

[6] There are certainly contexts in which it is meaningful to refer to validity with respect to test content, only, for a particular interpretation or use. Usually, however, it is test scores that are the focus of validity claims. The title of the ETS Report clearly indicates that the focus of the ETS Report is on test scores.

[7] Multistage testing complicates matters.

transformations of raw scores. Importantly, the values of coefficients (and most other statistics) are typically *different* for raw scores and scale scores that are non-linearly transformed raw scores.

Some Comments about Correlation Coefficients

Let *r* stand for the value of a correlation coefficient. Suppose *r* = .85. I am willing to admit that I have trouble understanding intuitively what *r* = .85 means, without supporting information. I can verbalize various mathematical statements about how *r* is defined,[8] and I can compute *r*, but that is far removed from having an intuitive understanding of the magnitude of a particular value of *r*. Summary statistics can help somewhat. A plot of scores for two variables is even more helpful to inform an interpretation of *r*, but plots are often not available. (None are provided in the ETS Report.)

Furthermore, consider two seemingly high values of *r* (say, .95 and .90) and two lower values of *r* (say .85 and .80). The arithmetic difference is the same (.05), but the meaning of the difference is definitely *not* the same.

Fisher once said, "the correlation coefficient … is an artificial concept (1925, p. 129)."[9] In his last published paper, Cronbach made a more nuanced comment when he said that "coefficients are a crude device that do not bring to the surface many subtleties (Cronbach, 2004, p. 394)." Strictly speaking, Cronbach made this comment in a discussion of reliability coefficients, but the comment applies equally (even more strongly, I would argue) to validity coefficients.

Again, I am not making an argument, here, against reporting correlation coefficients, per se. Coefficients have a long history in educational measurement, even though I am often skeptical of their inherent interpretative value.[10] I am strongly arguing, however, that the magnitudes of correlation coefficients per se are *not* a sufficient basis for recommending that ABA endorse or not endorse an "either/or" policy. Coefficients are simply too complex, and the interpretations of them typically obscure too many important matters. Whether or not coefficients are reported, to the extent possible, statistics should be reported that, as specifically as possible, quantify variability in different types of error such as those discussed previously.[11]

I believe the ETS Report might have been more informative if at least some plots of variables (along with descriptive statistics) were provided for a small representative sample of the 21 law

[8] For example, a correlation is the expected value of the product of z-scores for two variables.

[9] Tukey even went so far as to say, "most correlation coefficients should never be calculated (1954, p. 38)." In large part, I believe Tukey was arguing that examining plots of variables should take precedence over simply reporting correlation coefficients.

[10] Correlation coefficients are often convenient summary statistics in the sense that they can be used in conjunction with other statistics for the computation of certain much more easily interpreted statistics such as the slope of a linear regression.

[11] Personally, I would not recommend making an important decision based solely on the magnitude of a correlation coefficient; nor would I specify a particular value of a correlation coefficient as being indicative of a "valid test" or "valid scores."

schools (unidentified, of course). In no sense, however, am I arguing against reporting results over all schools and for schools with differing degrees of selectivity, as is done in the ETS Report. I am simply highlighting the obvious fact that admission is the responsibility of individual law schools, which means it is the individual law schools that must be able to defend the process and data used for admission.

Standard Errors of Measurement

A reliability coefficient is the ratio of "true score" variance to itself plus "relative" error variance. The square root of relative error variance is usually called a "standard error of measurement" (SEM), which quantifies how much measurement error (of a particular type) there is in test scores. Unfortunately, there is little reference in the ETS Report to SEMs. In addition, it would be highly desirable to provide (in a graph, for example) conditional SEMs (i.e., CSEMs) that quantify the errors for examinees with particular scores.[12]

Haertel (2006) provides probably the best currently available summary of various approaches to conceptualizing and estimating SEMs and CSEMs. Brennan (2001) discusses these matters from the perspective of Generalizability Theory. (Holland & Dorans, 2006, as well as Yen & Fitzpatrick, 2006, are partly relevant as well.) Note, in particular, that it is SEMs and CSEMs for *scale scores* that really matter here. Scale score CSEMS (based on non-linear transformations of raw scores), which are referenced by Haertel (2006, p. 84) and treated specifically by Kolen, Hanson, and Brennan (1992), as well as Kolen, Zeng, and Hanson (1996). See, also, Brennan (2021).

CSEMs for scale scores are relatively easy to explain (although challenging to estimate) and quite relevant here. They address the question, "If an examinee tested repeatedly (ideally with *different* forms of a test but the *same* administration conditions), how much variability in scale scores might be expected?" The answer to this question (even an imperfect answer) seems particularly relevant for a law school applicant as well as anyone evaluating the quality of a prediction system for law school admissions.

Comments Related to ABA Standard 503

ABA Standard 503 states that "A law school that uses an admission test other than the Law School Admission Test sponsored by the Law School Admission Council shall demonstrate that such other test is a valid and reliable test to assist the school in assessing an applicant's capability to satisfactorily complete the school's program of legal education."

With all due respect to the ABA, I believe that reference to a "valid and reliable" *test* in Standard 503 is just as misleading there as it is anywhere else. It is test *scores* that are reliable to some degree, and almost always it is test *score*s that possess different degrees of validity for different uses and interpretations. As noted previously, in my judgment this matter is *not* merely a terminological quibble.

[12] A CSEM for an examinee is the square root of "so-called" absolute error variance for that examinee.

It should be noted as well, that the Standard 503 reference to "satisfactorily complete the school's program of legal education" surely raises the prospect that at least one potential criterion for success is or should be a degree [which is a binary (0,1) variable] or the associated LGPA upon the attainment of a degree. I am not arguing whether or not either alternative should be adopted. That is an ABA policy issue. This matter, however, warrants clarification as it relates to the "either/or" decision. The ETS Report cites a paper by Wightman (2000) in support of overlooking this matter. It is unclear to me whether or not that paper is compelling, especially since it is over 20 years old.

Prediction, Linking, Equating, Concordance and Related Matters

Concordance is not discussed in most of the measurement literature. Undoubtedly, that is one reason why the topic is easily misunderstood. In particular, persons with little knowledge of concordance sometimes have uninformed beliefs about what scores should (or should not) be concorded, how concordances should be computed and interpreted, and how concordances should be evaluated. It is not surprising, therefore, that some persons harbor interpretations of concordances that are misleading or misinformed. Still, it is difficult---perhaps impossible---to overstate the centrality of concordance considerations (in both theory and practice) for an "either/or" decision by the ABA.

My explicit intent here is to explain concordance so that the ABA can make an informed "either/or" decision. My discussion may be viewed by some as needlessly arcane or excessive; others (particularly some experienced psychometricians) might even view my discussion to be patronizing. Neither view is my intent. However, I would much rather accept such criticism than fail to bring important concordance matters to the attention of the ABA. Importantly, note that, unless otherwise stated, all reference to "scores" (particularly "concorded scores") in this section refers to *scale* scores, not raw scores or number-of–items-correct scores.

An initial version of CASMA's report noted the importance of concordance if both the LSAT and the GRE are approved for use in the law-school admissions process under an "either/or" policy. CASMA noted that the ETS Report did not consider this matter. ETS responded that the matter was addressed. My search of the entire ETS Report revealed, however, that the word "concordance" is never used, which is entirely appropriate because nothing in the ETS Report, as far as I can tell, involves a concordance, as that term is typically used in practice and in extant literature. Later, in a memo entitled "*GRE-LSAT Concordance*" (dated July 5, 2021), ETS argued that the procedure in the memo is a concordance. I disagree, for the reasons explained in considerable detail next.

Two well-known current references that consider concordance in a manner that distinguishes it from other psychometric procedures are Holland and Dorans (2006, particularly pp. 193 and 217) and Kolen and Brennan (2014, chap. 10, particularly pp. 487-491). Kolen and Brennan (2014) is the most extensive, currently available, integrated treatment of equating, scaling and linking. (Concordance is a type of linking.) Holland and Dorans (2006) is an excellent, extensive chapter on linking and equating in the 4th edition of *Educational Measurement*. Holland and

Dorans are both former (and distinguished) ETS employees. *Educational Measurement* (Brennan, 2006) is widely viewed as the "bible" in the field of educational measurement. Also, Dorans, Pommerich, and Holland (2007), as well as von Davier (2010), provide edited books of chapters by various authors on the subjects of linking and related matters.

In general, various parts of this Addendum adhere to concordance matters as they are discussed in Chapter 10 of Kolen and Brennan (2014). To the best of my knowledge, there is no serious, substantive difference in perspectives between that reference and Holland and Dorans (2006) with respect to the matters discussed here.

The glossary of the *Standards for Educational and Psychological Testing* (2014) defines concordance as follows:

> **concordance:** In linking test scores for tests that measure *similar* (emphasis added) constructs, the process of relating a score on one test to a score on another, so that the scores have the same relative meaning for a group of test takers.

This definition is consistent with the use of the term "concordance" by Holland and Dorans (2006) and by Kolen and Brennan (2014). Note especially that this definition refers to "similar" constructs. The glossary of the *Standards* defines "construct" as "the concept or characteristic that a test is designed to measure." The word "construct" is commonly used in psychological testing. In educational testing, the word "content" is more frequently used. Here, I often refer to "similar content" because, for tests used in educational contexts, what is tested is typically discussed in terms of "content" (e.g., a table of content specifications for forms of a test).

There is no standard in the *Standards for Educational and Psychological Testing* (2014) that specifically mentions "concordance" per se. The commentaries for a few standards provide occasional reference to concordance, largely in the sense of general caveats about "linking" matters, but the 2014 *Standards* provide virtually no specific guidance tailored to the "either/or" context discussed here.

The next two sub-sections consider: (a) some important terminological matters that help to "frame" the meaning of concordance; and (b) a procedure proposed in the ETS Report that the Report authors seem to view as being a concordance. (As stated previously, the Report itself does not use the word "concordance," but a related ETS memo does.[13]) Following these two sub-sections, concordance and population dependence are treated in more detail.

<u>Fundamental Terminological Distinctions</u>

To explain "concordance," it is definitely helpful, and I think likely essential, to have at least some understanding of the distinctions among "prediction," "linking," and "equating."

- *Prediction* applies when scores on one variable, say GRE-V, are used to make a certain kind of statistical estimate of scores on another variable, say LGPA. Prediction is

[13] When this Addendum was being written, CASMA received a confidential memo from ETS entitled "GRE-LSAT Concordance" and dated July 5, 2021.

directional in the sense that the statistical relationship for predicting LGPA scores from GRE-V scores is *not* the same as the statistical relationship for predicting GRE-V scores from LGPA scores. For prediction, scores on both variables must be available for the *same* persons. Most applications of prediction in measurement are linear (i.e., a straight line relationship). Sometimes, two or more predictor variables are used. Importantly, a prediction relationship is *population dependent*; i.e., prediction generally differs for different groups.

- *Linking*, by contrast, is usually a *non-linear* relationship created in such a way that the relationship is *bi-directional*. So, for example, if a score of 150 on GRE-V were "linked" to a score of 160 on LSAT, it is equally defensible to say that a score of 160 on LSAT is linked to a score of 150 on GRE-V. Stated more technically, if a GRE-V score of 150 is linked to a LSAT score of 160, that means then the proportion of examinees (in some specific group) who got 150 or less on GRE-V is the same proportion (in the same group) as got 160 or less on LSAT. Importantly, a linking relationship is *population dependent*; i.e., the relationship generally differs for different groups, except as noted next.
- *Equating* is a specific type of linking. As such, equating is *bi-directional* in the sense discussed in the previous "linking" bullet. Importantly, the two variables involved in an equating are scores for "forms" of the *same* test, *with the forms developed according to the same content and statistical specifications and administered under the same conditions.* Under these very specific circumstances, it is generally acknowledged that equating is typically *population invariant*; i.e., equating results are generally quite similar for different populations (e.g., males and females). In this sense, equating satisfies a substantially higher "bar" than either prediction or linking. I believe there is good reason to assert that scores on forms of the GRE and LSAT tests are likely well-equated, but that definitely does *no*t guarantee that a relationship *between* LSAT and GRE-V scores is necessarily population invariant.

Equating is largely unmentioned in the ETS Report. If equating is successful for all forms of the GRE and LSAT tests (which is an assumption I am willing to make, here), it is not necessary to consider equating issues in the context of concordance. It is worth noting, however, that any realistic data set used to obtain concordances almost certainly will involve scores from a large number of *different forms* of both the GRE and the LSAT tests. Also, if the test specifications (and/or administration modes) for the GRE and/or LSAT tests change over time (even if the names of the tests remain the same), there likely will be a need for new concordances.

<u>The ETS GRE-LSAT Procedure</u>

Both Holland and Dorans (2006) and Kolen and Brennan (2014) are very careful to distinguish among terms such as "prediction," "linking," "equating," and "concordance." The ETS report is often rather unclear about such distinctions. To many readers of the ETS Report, these distinctions may seem arcane and unimportant. Here, however, I believe these distinctions are likely to be crucial for any ABA decision about an "either/or" policy.

It is not CASMA's contractual responsibility to examine all ETS (or other) research that involves the GRE. However, the ETS claim that their procedure is a concordance was so unexpected to

me that I did a "deeper dive" into the matter. As best I can determine, what ETS did was propose a procedure (not properly called a concordance, in my judgment) that relates LSAT scores and a weighted combination of GRE-V and GRE-Q scores. To do so, ETS used the exact same data (21 schools with a total of 1,587 students) referenced and used in the ETS Report.[14] I will call this the "ETS Procedure."

Basic Steps. It is my understanding that the basic steps used in the ETS Procedure are as follows:

1. Use multiple linear regression to obtain a *predicted* LSAT score given GRE-V and GRE-Q scores. The nature of multiple regression is that it determines statistically "optimal" weights for the predictors, GRE-V and GRE-Q in this case. These weights turned out to be approximately .6 for GRE-V and .4 for GRE-Q.
2. Conduct an equipercentile "linking" of the actual LSAT scores and the predicted LSAT scores resulting from Step 1. (Step 2 is *not* an equating.)

Clearly, the ETS Procedure has both prediction and linking features. Specifically, the ETS Procedure is an adjusted linear prediction, with the adjustment taking the form of an equipercentile linking. As such, the ETS Procedure is *not* a concordance, as that term is generally used in practice and in extant literature cited previously. Note that the ETS "marries" a *one-directional* prediction and a *bi-directional* linking. To the best of my knowledge, there is no precedent for such a procedure, and no clear interpretation of its psychometric characteristics.

Admittedly, however, the ETS Procedure has an appearance similar to a typical concordance, and, as such, the procedure may be sensible for some purposes that I judge to be outside the purview of this Addendum.[15] As far as I know, the ETS Procedure has not been used or studied elsewhere in published literature, except for the ETS "Comparison Tool" and "Questions" document cited in footnote 15.

Data. The data ETS used to perform the two-step procedure outlined above is described on page 4 of the ETS Report. All 21 schools had LSAT scores, as well as UGPA and LGPA, for their matriculated students. Ten institutions also had GRE scores for (most?) matriculated students. For the other 11 schools (whose matriculated students did not have GRE scores) a special study was conducted to obtain GRE scores for as many law school students as possible. Clearly, ETS deserves credit for making serious efforts to obtain GRE scores for these students.

The data are limited, however, in three senses. First, the matriculated-students sample size (1,587) is relatively small. Second, the GRE data for the 11 special study schools are clearly

[14] These students matriculated between the 2010-2011 and 2016-2017 academic years, which means that at least some of the data are rather old.

[15] ETS provides a web-based Excel worksheet entitled "Comparison Tool for Law Schools – 2018" that uses GRE-V and GRE-Q to provide a "Predicted LSAT Score." Note the word "predicted" as opposed to "concorded." Some information about this tool, and the ETS Procedure, are available on the web in a brief document entitled "*Frequently Asked Questions about the GRE Comparison Tool for Law Schools (March 2018)*." Except for this document, I was unable to find any other formal ETS document (e.g., a Research Report) that provides psychometric justification for this "Comparison Tool" or calling it a concordance.

atypical in terms of the data collection procedure and, therefore, perhaps not as credible as the data for the other 10 schools. Third, and perhaps most importantly, the data used by ETS are for *matriculated students, only*, not applicants. Consequently, we do not know if the ETS GRE/LSAT relationship for matriculated students applies to non-admitted students. (This is a type of possible "selection error" mentioned previously.) A fundamental issue is that a prediction system is used with applicants, not just a subset of those who matriculate.[16]

Other Comments. The ETS Procedure specifically uses a weighted combination of both GRE-V and GRE-Q, although clearly the basic approach could involve GRE-V, only. To simplify the verbal discussion of various issues, much of the subsequent discussion of the ETS Procedure refers to use of GRE-V, only. Joint use of the weighted composite of GRE-V and GRE-Q is explicitly considered when it is necessary to do so.

Admittedly, the above comments about the ETS Procedure beg questions such as: "What is a concordance?"; "How is a concordance typically obtained?"; "What are the statistical features of a concordance?"; and "Why do these features matter?" These questions are considered next, without addressing all relevant details. I am *not* claiming that certain ETS staff are unaware of these matters. I am claiming, however, that these matters are largely not addressed in the ETS Report, and an understanding of them is important for the ABA to make an informed "either/or" policy decision. (An ETS confidential memo dated July 5, 2021, calls the ETS Procedure a concordance but, for reasons discussed next, the ETS Procedure is not a concordance.)

Concordance, Content Similarity, and Population Invariance: Basic Ideas

A concordance is definitely neither a prediction nor an equating in the above senses of the terms. Rather, concordance is a particular type of linking and is, therefore, *bi-directional*. Importantly, note that concordance is a term that is reserved for tests that are intended to measure ***similar constructs*** under similar measurement conditions. See, for example, Holland and Dorans (2006, p. 193), Kolen and Brennan (2014, p. 499), and the previously cited definition of "concordance" in the *Standards for Educational and Psychological Testing* (2014). An essential defense of a concordance involves sound evidence of *content similarity* for the tests (e.g., LSAT and GRE, or part of GRE) whose scores are concorded. Secondarily, but importantly, a concordance claim is supported if there is relatively little population dependence for relevant subgroups.

Content Similarity. Subject matter experts are used to inform judgments about construct similarity. Importantly, no statistical or psychometric procedure, in and of itself, can lay claim to being a "concordance" without a "similar content" claim being supported in some manner and to some degree. Also, if content similarity claims are clearly suspect, then the word "concordance" does not apply to the linking relationship.

It follows that the scores that result from applying the ETS Procedure, which involves both GRE-V and GRE-Q, cannot lay claim to being concorded with LSAT scores unless (at a

[16] Data for concordances are often obtained by matching data sets from testing companies, rather than using data for matriculated students at colleges, universities, etc. Testing companies have test scores and identifying information for students who chose to take their test, but, to the best of my knowledge, testing companies typically do not know the admission status for all applicants.

minimum) it can be argued that the *presence* of GRE-Q in the ETS Procedure, and the *lack* of a math test in LSAT, is a content *irrelevant* matter. I have some credentials in math, but none of consequence for the content in GRE-V or LSAT. Still, it seems much more likely to me that LSAT and GRE-V are more similar in content than are LSAT and the content represented by the GRE composite (.6 GRE-V + .4 GRE-Q).

Population Invariance. Suppose an equipercentile linking of LSAT and GRE-V scores is obtained for males, and a separate equipercentile linking is obtained for females. If the differences between the two equipercentile linkings are small, then it could be claimed that the concordance was successful or useful. That is, we would expect pretty much the *same admissions decisions* (based on predicted LGPA) for males whether LSAT scores were used directly in the prediction equation, or the concorded GRE-V scores were used in place of the LSAT scores in the prediction equation; and a similar statement would hold for females.

By contrast, if the equipercentile linkings for males and females were substantially different, then admissions decisions for some males would be different depending on whether LSAT scores were used or the concorded GRE-V scores were used, and a similar statement would hold for females. If so, the concordance would be problematic.[17] In short, studying population invariance is a psychometric method for informing a decision about the utility of a concordance in situations such as the LSAT/GRE "either/or" decision.

There are usually practical limitations to studying subgroup linkings, including concordances. For example, equipercentile linking procedures typically require at least 3,000 examinees, which may be unattainable for certain subgroups. Sometimes, as few as 1,500 examinees may be adequate. For smaller sample sizes, linear linking procedures can be used, but they are not nearly as good for studying subgroup dependency throughout the entire score scale. The most challenging scenario would occur when, for whatever reason, subgroup linking procedures are not performed. In such a case, judgments by content matter experts would have to carry the entire weight for a decision about content similarity, with an empirically unsupported assumption about subgroup invariance.

In the context of a data-based example (somewhat artificial), Chapter 10 of Kolen and Brennan (2014) discuss and illustrate the use of various statistics and visual displays of data that can be employed to examine subgroup invariance. Kolen and Brennan (2014) also provide additional references relevant to concordance and other types of linking.

ACT-SAT Concordances: An Old Example

To ground further discussion of concordance, I think it useful to provide an overview of an old example of ACT-SAT concordances.[18] I am not proposing ACT-SAT concordances as a model

[17] The word "concorded" has both a "process" connotation and an "outcome" connotation. Typically, in my experience, that has not been a source of confusion. In the LSAT/GRE-V context, however, this may be causing confusion.

[18] I have had differing degrees of involvement in the various ACT-SAT concordances that have been produced over the last several decades, when I was either an ACT employee or a College Board consultant, but I never had primary responsibility for conducting these concordance studies.

for how concordances must be obtained, but the issues addressed in discussing them are relevant to the general topic of concordance for present purposes.

Nothing I say here should be viewed as an evaluative judgment about either testing company (ACT or the College Board, which owns the SAT), either testing program, or the ACT-SAT concorded scores per se. In addition, the discussion here is substantially simplified and based in part on my long-term memory of some complicated matters. Finally, the following discussion is intended to illustrate issues, only, without any claim that the ACT-SAT concordances are a definitive example of exactly how concordances must be obtained.

Note that, to the best of my knowledge, since the late 1980's (and perhaps earlier), there never has been a need for an ACT/SAT "either/or" *policy* decision, per se, because virtually every college/university in the US had a prediction system based on *either* the ACT *or* the SAT, and many institutions accepted either the ACT or the SAT. Consequently, both ACT and the College Board recognized the need for concordance tables in order to serve colleges and universities, as well as to maintain and/or enhance their competitive position in the marketplace.

The ACT Assessment has had four tests ever since its inception in 1959, although the actual test specifications (and, to an extent, the test names) have changed over time. Here, attention is given exclusively to the version of the ACT Assessment introduced in the late 1980s. The four tests were English (E), Math (M), Reading (R), and Science (S). Each test reported scores on a scale of 1, 2, …, 36. There was also a reported composite score (C) for each examinee that was obtained by averaging each examinee's four scale scores and rounding to an integer, which implies that the composite scores were also 1, 2, …, 36.

Over past decades, the SAT has had changes in both the number of tests, the names of tests, and the test specifications.[19] Here, attention is given exclusively to the "old" SAT that had only two primary tests: SAT-V and SAT-Q. Reported scale scores for both tests were 200, 210,…, 800; i.e., for both tests there were 61 possible reported scores for an examinee. For most of the time that this version of the SAT was in effect, the College Board did not report a composite score to examinees. It was common practice, however, for others (e.g., colleges) to compute a composite that ranged from 400 to 1600 for both tests.

In total, then, there were five ACT scores and three SAT scores (including the composite). Statistically, any combination of ACT scores could have been linked with any combination of SAT scores, but that does not mean that all of these possible linkages would have been meaningful or defensible concordances. (Recall previous discussions of content/construct issues.)

The first step was to obtain the data needed for the concordances. That was a complex issue, because both ACT and the College Board wanted to protect the privacy of their own data sets. My recollection is that a third party (unknown to me) facilitated obtaining a final merged data set

[19] The current SAT introduced about five years ago is dramatically different from previous versions. Information about concordances for the current versions of the ACT and SAT can be found at https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf. In my opinion, this document provides particularly good, brief discussions of important concordance matters.

(devoid of identifying information) that contained the examinee scores and some demographic variables necessary to obtain concordances. (This is a substantially simplified explanation based on old memory that may be flawed.) In the end, the matched sample involved tens of thousands of examinees.

The second step was for the two companies to agree on which tests (or combination of tests) would be concorded. My recollection is that it was relatively easy for the two testing companies to reach agreement about concording the composites (see actual results for doing so in Kolen & Brennan, 2014, pp. 487-491) and concording the two math scores. It was more difficult to reach agreement with respect to the scores to be included in other concordances.

The third step was to obtain the concordances. This was done by jointly through a somewhat complex arrangement. Importantly, in the end,

- there was *one and only one* reported concordance table for each pair of concorded scores[20],
- each concordance was based on nearly all available data, and
- *both* ACT and the College Board endorsed use of *each* concordance.

The second and third steps were potentially iterative in that the final decision about which tests to concord could have been informed, in part, by the analyses in the third step.

Some Implications for the "Either/or" ABA Policy Decision

The above three steps for the ACT-SAT concordances have been discussed not because every concordance situation must follow the exact same steps. The basic message is that the issues involved in these three steps are challenging and typically must be addressed in some manner if the ABA chooses to pursue an "either/or" policy decision. I believe that doing so likely will require the participation or input, in some sense, of both ETS and LSAC, since I assume each of them claim some degree of ownership of, or responsibility for, the scores on their respective tests. Also, neither ETS nor LSAC likely would have direct access to the applicant scores for both testing programs for all schools.

In at least one sense, the GRE and LSAT concordance issues are considerably simpler than the ACT-SAT issues. Specifically, there are fewer possible GRE test combinations that might be considered for concordance with LSAT. For discussion purposes here, I assume that there are only two concordances that are likely to receive serious consideration: (i) GRE-V scores concorded to LSAT scores; and (ii) a combination of GRE-V and GRE-M scores concorded to LSAT scores.[21] Among other things, the next section reviews reasons why (ii) may not be an easily defended concordance.

[20] In theory, one element of a pair could itself be a sum or average of selected ACT scores.

[21] This does not mean that I am arguing against attempting to include GRE-AW in a concordance, although I doubt that doing so would be very helpful or successful. Of course, GRE-AW might serve an auxiliary role in admission decisions.

Content and Related Considerations Revisited

With respect to (ii), above, as noted previously, I have serious reservations about jointly including GRE-V and GRE-Q in a concordance with LSAT. The incorporation of GRE-Q clearly seems to me to violate the basic notion of content similarity in concordance.

Even if I were to overlook that concern, however, another matter deserves attention. Specifically, recall that the ETS procedure employs a statistical weighting for GRE-V (about .6) and GRE-Q (about .4). Now, suppose an attempt was made to concord scores for .6 (GRE-V) + .4 (GRE-Q) with LSAT scores. The (approximate) .6 and .4 weights are definitely population dependent, since they are based on prediction. However, it seems likely (at least plausible) that such weights would be interpreted by institutions and applicants as having an explicit content-based interpretation as defined by experts---e.g., something like, "With respect to content, the GRE-Q is judged to be .4/.6 = 2/3 as important as GRE-V for success in law school." If so, the .6 and .4 weights will be misleading, because obtaining judgmental content-based weightings is not the same process as obtaining statistical optimal weightings.

In short, although it is not psychometrically wrong per se to use different weights in forming a score that is part of a concordance, I do not recall ever seeing that done before. Doing so is likely to be misleading, unless the weights have a coherent content defense. This is clearly a psychometric issue, but it is also an "appearance" issue, and appearances can matter when public defenses are required. For example, besides the fact that they are unequal, the (approximate) .6 and .4 weights suggest a degree of precision that content matter experts probably would have difficulty defending.

Again, with respect to concordance, perhaps the most crucial consideration is the degree of similarity in content *between* the two programs' tests whose scores are concorded. All other things being equal, it is expected that a concordance will be *less* population dependent when the content is *similar* for the tests. In addition, in such cases, the concordance is likely to be more defensible and more understandable to various stakeholders.

My concern here (and indeed throughout this Addendum) is emphatically *not* the quality of any ETS or LSAC test (or section of a test), as judged by content matter experts. I assume that both ETS and LSAC have highly competent content matter expects who deal with such issues. Rather, my concern is with content/construct similarity *across* testing programs.

In summary, I assert, that content considerations, in the sense discussed above, are very important for concordance issues and need be addressed somehow. While I have reservations about using GRE-Q scores in a concordance with LSAT scores, I take no position with respect to the appropriateness (or lack thereof) of using GRE-Q (or any well-constructed math test) in some other part of the law-school admissions process.

Recommendations

I do not believe that the ETS Report is a sufficient basis for the ABA to adopt an "either/or" policy involving the LSAT and GRE. Emphatically, this judgment should not be interpreted in any way as a criticism of the GRE or LSAT, per se. Rather, this judgment is based on the *proposed use* of the GRE as described in the ETS Report.

As discussed previously in this Addendum, my two major concerns are as follows.

1. The Report has technical limitations that, in my opinion, preclude it from adequately informing an "either/or" policy and/or supporting the implementation of such a policy in a psychometrically defensible manner. (Of course, the Report may be quite adequate for other purposes.)
2. The adequacy of the data upon which the Report relies is suspect, I believe. One issue is that the conditions under which much of the GRE data were obtained often do not faithfully mirror the conditions under which GRE scores would be obtained and used in an "either/or" environment. Also, the data set is probably too small to provide an adequate empirical basis for judgments about concordance under an "either/or" policy. In addition, it is important to note that the data are for matriculated students, only. Matriculated students are clearly relevant, but so are unadmitted applicants.

While I have reservations about the current study, I think there is a reasonable possibility that a different study could be conducted that might well support an "either/or" decision by the ABA. For that reason, I would suggest that the ABA consider supporting and/or encouraging such a study, which I will call a Pilot Study, here.

I suggest that the following two basic principles guide the design and implementation of a Pilot Study.

- There need to be a sufficiently large number of applicants (both admitted and not admitted) who take both the LSAT and the GRE.[22] There likely will need to be some motivation (money?) for some/many applicants to take both tests.
- The conditions under which the LSAT and GRE are taken by applicants (which includes both matriculated and non-admitted examinees) should be reasonably similar. For example, it is probably not advisable to mix quite old data from one testing program with new data from the other program.

It might be reasonable to include some data from the ETS study. I cannot make that statement definitively because I do not have a detailed enough understanding of the data.

I think the Pilot Study could be completed in a year's time or less. Here, I do not consider the obvious important matter of cost. Provided next is a discussion of three issues (sample size,

[22] I recognize that there are already *some* law schools that allow applicants to take either the LSAT or the GRE, but I doubt that there are a sufficient number of applicants for the Pitot Study. Even if there were enough applicants, the current "either/or" schools are self-selected, which means it cannot be assumed a priori that they faithfully represent all law schools.

analyses, and process) that merit consideration. My discussion of these matters is intended to be helpful or informative, but by no means definitive.

Sample size

There is no magic number that designates how large the sample size must be, but I think there are some reasonably obvious considerations.

- All other things being equal, a larger sample size is preferred over a smaller sample size.
- In a sense, the quintessential concordance would be an equating, in which the content of the two tests is undeniably similar. The best equating, in my opinion, uses equipercentile procedures (EP) with a randomly equivalent groups (RG) design---call this EP/RG. In my experience EP/RG with 3,000 or more examinees is both typical and adequate. Sometimes EP/RG appears to be satisfactory with as few as 1,500 examinees, but that is unusual.
- Studying population dependence of concordance requires using EP with subsets of the data (e.g., males, females, and perhaps various racial/ethnic groups if sample sizes permit), which obviously means smaller sample sizes are available for studying subgroup invariance than for an overall-persons concordance.
- Clearly, there is an upper limit to how large the sample size could be given the number of applicants in a given year, as discussed next.

From a search of the ABA website, my understanding is that there are 205 ABA approved law schools that confer a J.D. degree, with a total enrollment of a little less than 115,000. Assuming a three-year program, that means that the average number of first-year students (over all schools) is about 38,333.

I do not know how many applicants any school may have, but suppose there are, on average, *Y* times as many applicants as admitted students. Also, suppose that for Pilot Study purposes, *Z* is the proportion of applicants who take both LSAT and the GRE. If so, letting "x" stand for multiplication, the matched data for establishing a concordance would consist of

$$[Y \times (115{,}000/3)] \times Z = (Y \times Z) \times 38{,}333 \text{ examinees.}$$

For example,

- if *Y* were 1.5, and *Z* were 1/10, then the matched data set would consist of about 5,750 examinees; and
- if *Y* were 2, and *Z* were 1/10, then the matched data set would consist of about 7,667 examinees.

Assuming *Y* and *Z* can be approximated with reasonable guesses, this type of thinking puts an upper limit on how large the sample size conceivably "could be" for the Pilot Study.

The "would like" sample size obviously must be less than the "could be" sample size. I definitely do not have enough information to specify the "could be" sample size. At this point, given my current state of knowledge/ignorance of many relevant matters, on balance, I think that perhaps

6,000 might serve as a reasonable minimum "would like" sample size, but that probably would not permit as deep a study of population dependence for racial/ethnic groups as for gender.[23] If a sample size of 6,000 is judged to be unattainable, then 3,000 might be sufficient for some analyses. It is difficult for me to believe that a sample size of less than 3,000 would be adequate for equipercentile subgroup analyses.

There are, of course, other ways to think about obtaining an adequate sample. For example, the study design could focus on a "census" testing of all (or most) applicants in a stratified sample of schools based on selectivity, which has similarities with what ETS discussed doing in their Report. Note, however, that serious attempts should be made to get *both* LSAT and GRE data for applicants, not just those students who are admitted.

Analyses

For each test-pair under consideration (e.g., LSAT and GRE-V, or LSAT along with a weighting of GRE-V and GRE-Q) content similarity analyses need to be conducted. Then, for any test-pair judged to be content similar, concordances should be obtained and, to the extent possible, population invariance should be studied. Based on what I currently know, I suspect that content considerations, as well as empirical analyses, *may* support concording LSAT and GRE-V. As noted previously, I am skeptical about a concordance of LSAT with GRE-V and GRE-Q jointly, but it could be examined as well.

Also, to the extent possible, supporting analyses (or at least discussions) should consider the sources of error discussed earlier in this Addendum. Correlation coefficients could be reported as judged helpful or necessary, but they are not sufficient. Plots of results for selected schools would be especially helpful, I think.

Process

I believe it is important to pay careful attention to the process whereby a Pilot Study would be conducted. The following comments may merit consideration.

If the various ACT-SAT concordance studies are used as precedent for how to proceed, then it may be important that the Pilot Study be planned and conducted with some degree of involvement by both ETS and LSAC. In a similar vein, if the ACT-SAT studies are viewed as precedent, it may be sensible to involve a third entity for certain specific tasks.

In any case, it seems to me that a Pilot Study will require some degree of active involvement by the ABA, since it is only the ABA that can make an "either/or" decision. In particular, it seems necessary that there be written plans for a Pilot Study. The plans should be detailed enough that the ABA feels confident that the study results would provide enough information for the ABA to make an informed decision.

[23] If sample sizes do not permit using equipercentile procedures for some concordances, linear procedures may be adequate for some uses (see Kolen & Brennan, 2014, chap. 10).

Concluding Comments

The score scales for GRE-V and LSAT are definitely *different*, although they have certain numerical characteristics that make them appear similar. The most notable similarity (in appearance) is the overlapping set of numbers that characterize the scale-score ranges: 120 to 180 for LSAT, and 130 to 170 for GRE-V. (Note, however, that the number of possible scale score points for LSAT is 61, while the number of possible scale score points for GRE-V is 41, which are quite different values.)

Even though the score scales for GRE-V and LSAT are *different*, as best I can determine, the reported *scale-score* SEMs for the two tests are *both* about 2.5 scale score points.[24] This almost certainly means that the scaling of the two tests used substantially different assumptions, procedures, and/or populations. That, in turn, might mean that concorded values for LSAT and GRE-V may not be as similar as some might expect, given the seemingly substantial overlap of scale-score ranges.

Also, it is reasonable to assume that the overlap in scale-score ranges may lead to confusion for some users of the concordance. Stated differently, when score scale ranges overlap, it can be challenging for some users to correctly distinguish between scores for the two tests. (For example, it is quite unlikely that an LSAT score of 160 has the same meaning or interpretation as a GRE score of 160.) This is not a psychometric limitation, per se, but it may confuse certain users of a concordance.

Importantly, if concordance is successful, the responsible parties (i.e., ETS, LSAC, and I assume ABA) will need to inform applicants that different concorded results do *not* mean that it is easier to get admitted using one test rather than the other.[25] This may be a challenging message to convey when the score scale ranges appear similar, but concorded scores are (or may be) noticeably different.

Finally, this Addendum began by framing the "either/or" decision in the context of the use of LSAT or GRE in a traditional linear-prediction statistical context (see section on "Context and Use: an Overview"). That focus was used to simplify certain discussions in this Addendum. Importantly, however, the issues involved in concording LSAT and GRE are not restricted to that context. I do not know the specific details of the process used by any law school to make admission decisions, which means that I do not know exactly where, when, or how test scores

[24] For both GRE-V and LSAT, finding the estimated SEM value of about 2.5 involved some non-trivial web searches. The two documents upon which I finally relied are "ETS GRE Reliability and Standard Error of Measurement" (undated, but likely very recent) and "LSAT-interpretative –guide/2019-2020." Of course, the statistical procedures and populations used to obtain these estimates are likely quite different (perhaps dramatically different). Note that, since LSAT has over twice as many raw score points as GRE-V, it is virtually certain that a proportion-correct *raw-score* SEM for LSAT would be considerably *smaller* that a proportion-correct *raw-score* SEM for GRE-V.

[25] For example, suppose an examinee believes s/he needs an LSAT score of at least 160 to have a good chance of being admitted to some particular law school. If an LSAT score of 165 is concorded to a GRE-V score of 155, that does *not* mean that it is easier for this examinee to get admitted using LSAT. The *Standards for Educational and Psychological Testing* (2014) is clear with respect to cautioning users about reasonably anticipated misuses of test scores (see especially Standard 5.3).

are part of the process. Nonetheless, the concordance issues involved in an "either/or" decision are still relevant as long as test scores are part of the process in some meaningful sense.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (Ed.) (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Brennan, R. L. (April 10, 2021). *Generalizability theory: history/contributions, extensions, and challenges*. Division D Linn Award Address at Annual Meeting of AERA.

College Board & ACT (2018). *Guide to the 2018 ACT/SAT concordance.* See: collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. (Editorial assistance provided by R. Shavelson.) *Educational and Psychological Measurement, 64,* 391--418.

Dorans, N. J., Pommerich M., & Holland, P. W. (Eds.) (2007). *Linking and aligning scores and scales.* New York: Springer-Verlag.

Fisher, R.A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: American Council on Education/Praeger.

Klieger, D. M., Bridgeman, B., Tannenbaum, R. J., Cline, F. A., & Olivera-Aguilar, M. (2018). *The validity of GRE General Test scores for predicting academic performance at U.S. law schools*. (Research Report No. ETS-RR-18-26). Princeton, NJ: Educational Testing Service.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd ed.)*. New York: Springer-Verlag.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29,* 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores Using IRT. *Journal of Educational Measurement, 33, 129-140.*

Tukey, J.W. (1954). Causation, regression and path analysis. In O. Kempthorne, T.A. Bancroft, J.W. Gowen & J.L. Lush (Eds.) *Statistics and mathematics in biology* (pp.35–66). Ames, IA: Iowa State College Press.

von Davier, A. A. (Ed.) (2010). *Statistical methods for test equating, scaling, and linking*. New York: Springer-Verlag.

Wightman, L. E. (2000). *Beyond FYA: Analysis of the utility of LSAT scores and UGPA for predicting academic success in law school* (Research Report No. 99-05). Newtown, PA: Law School Admission Council.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 11-154). Westport, CT: American Council on Education/Praeger.