Research Report

Validity, Reliability, and Fairness Evidence for the JD-Next Exam

ETS RR–24-04

Steven Holtzman
Jonathan Steinberg
Jonathan Weeks
Christopher Robertson
Jessica Findley
David Klieger

December 2024

ETS

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

ETS RESEARCH REPORT

Validity, Reliability, and Fairness Evidence for the JD-Next Exam

Steven Holtzman[1], Jonathan Steinberg[1], Jonathan Weeks[1], Christopher Robertson[2], Jessica Findley[3], & David Klieger[1]

1 ETS, Princeton, New Jersey, United States
2 Boston University, Boston, Massachusetts, United States
3 University of Arizona School of Law, Tucson, Arizona, United States

At a time when institutions of higher education are exploring alternatives to traditional admissions testing, institutions are also seeking to better support students and prepare them for academic success. Under such an engaged model, one may seek to measure not just the accumulated knowledge and skills that students would bring to a new academic program but also their ability to grow and learn through the academic program. To help prepare students for law school before they matriculate, the JD-Next is a fully online, noncredit, 7- to 10-week course to train potential juris doctor students in case reading and analysis skills. This study builds on the work presented for previous JD-Next cohorts by introducing new scoring and reliability estimation methodologies based on a recent redesign of the assessment for the 2021 cohort, and it presents updated validity and fairness findings using first-year grades, rather than merely first-semester grades as in prior cohorts. Results support the claim that the JD-Next exam is reliable and valid for predicting law school success, providing a statistically significant increase in predictive power over baseline models, including entrance exam scores and grade point averages. In terms of fairness across racial and ethnic groups, smaller score disparities are found with JD-Next than with traditional admissions assessments, and the assessment is shown to be equally predictive for students from underrepresented minority groups and for first-generation students. These findings, in conjunction with those from previous research, support the use of the JD-Next exam for both preparing and admitting future law school students.

Keywords Law school; admissions; fairness; validity

doi:10.1002/ets2.12378

U.S. law schools need reliable, valid, and unbiased tools to predict whether applicants are likely to be successful in their programs of legal education for the juris doctor (JD) degree. Law schools have historically relied on the Law School Admission Test (LSAT) and, more recently, the *GRE*® examination, which are general tests of language ability, analytical reasoning ability, and, in the case of the GRE, basic quantitative reasoning ability (Klieger et al., 2016). This approach is typical, where most standardized tests assess a combination of language ability, quantitative ability, writing ability, and analytical reasoning ability (Kuncel & Hezlett, 2007).

Nonetheless, a range of other approaches have been developed. One approach—called *proximal*, *trial studying*, or *curriculum sampling*—uses a test that seeks to mimic representative parts of an academic program, reducing the inferential gap between test and ultimate performance (Farley et al., 2019; Niessen et al., 2016, 2018). Another approach, called *dynamic* or *learning potential*, seeks to measure a student's development through the course of a learning experience (Lidz, 1995; Sternberg & Grigorenko, 2020). Although they have roots in the 1920s and are more widely used in Europe, these alternative approaches have only recently been applied to and rigorously evaluated in the context of U.S. legal education (Findley et al., 2023; Shultz & Zedeck, 2012).

One feature of these alternative approaches is their design for a specific discipline, such as chemistry or law, so that the test measures performance in the same domain where the test aims to predict performance, reducing the inferential gap. Accordingly, in a systematic review of the literature, authors noted that "the strongest predictors are tests with content specifically linked to the discipline" (Kuncel & Hezlett, 2007, p. 1080). The discipline specificity of such exams makes the approach similar to "work sample" testing, used in the industrial–organizational psychology field with noted success (Den Hartigh et al., 2018; Roth et al., 2005; Schmidt & Hunter, 1998; Tucker, 2016).

Corresponding author: J. Weeks, E-mail: jweeks@ets.org

A second feature of some of these alternative approaches is pairing the test with a supportive learning environment so that a learner can be fully prepared for the exam by participating in the keyed course. In this way, "what is tested is not just previously acquired skills, but the capacity to master, apply, and reapply skills taught in the dynamic testing situation" (Sternberg & Grigorenko, 2020, p. 29). Given that law schools are in the business of changing students, helping them grow into successful legal professionals, admissions officers may be particularly interested not only in the abilities a student has at the beginning of law school but also in the student's ability to grow through law school.

In theory, a proximal, dynamic approach to testing could reduce racial, ethnic, and class disparities in test scores. Extant admissions tests, such as the LSAT, are known to have large score disparities, with Black test takers having mean scores about 2 standard deviations (or 11 points) lower than White non-Hispanic test takers, for example (Dalessandro et al., 2014; Lauth & Sweeney, 2022). Some argue that these test score gaps reflect background disparities in wealth and educational opportunity, which are themselves "an enduring legacy of past inequity" (Taylor, 2014, p. 16). As such, they center as "merit" certain cognitive assets that are acquired through accumulated privilege (Guinier, 2015). The social distribution of learning potential may be more uniform than the distribution of scores on general standardized tests. Indeed, research has shown that racial and ethnic gaps on the LSAT are larger than differences in undergraduate grades, law school grades, or measures of subsequent success in the legal profession (Kidder, 2001). A systematic effort to measure learning potential in a supportive environment may thus reduce racial disparities in scores and support racial inclusion in programs of education.

While general disparities in backgrounds are problematic, a more specific and proximate mechanism for racial, ethnic, and class disparities in standardized test scores could be found in disparate levels of access to and use of test preparation programs, which can be expensive in terms of time and money (Amabebe, 2020). Coached test preparation has been shown to have a benefit of about 25% of a standard deviation on standardized test scores (Kuncel & Hezlett, 2007). Major test providers, such as the Law School Admissions Council (LSAC) and ETS, offer some free test prep programs but do not actually design their tests to assess learning in a specific, free course of education.

Building on these insights, the JD-Next program is a fully online, noncredit, 7- to 10-week course to train potential JD students in case reading and analysis skills, prior to their first year of law school (Findley et al., 2023). The JD-Next exam is designed to measure learning potential, with multiple-choice (MC) and essay questions aligned with the learning objectives of the JD-Next course, which covers classic cases in contract law doctrine as a foundation for developing essential skills of reading and analysis that could be applied to any law school class. More specifically, the learning objectives assessed by the exam include writing an effective legal analysis of a hypothetical fact pattern; identifying and articulating the legal issue in a judicial opinion; identifying and articulating the dispositive facts of a judicial opinion; identifying and synthesizing the rule of law as applied in a judicial opinion; distinguishing the legal reasoning of the plaintiff, defendant, and court in a judicial opinion; identifying and articulating the procedural posture of a given judicial opinion; and identifying and articulating the holding and disposition of the court in a given judicial opinion. To hone these skills, students learn from classic contracts cases, for example, *Hamer v. Sidway* (1891), which involves the distinction between a contract and a mere promise to make a gift and *Ever-Tite Roofing Corp. v. Green* (1955), which teaches the concept of acceptance by performance.

The JD-Next program and exam have two purposes: They can be used to support students' readiness for law school by teaching critical skills in case briefing and legal analysis ahead of law school, or they can serve as a supplemental admissions test for predicting students' ability to be successful in law school. Because there are two purposes, both undergraduate and admitted students may find it beneficial (Buzick et al., 2023; Cheng et al., in press). Undergraduate students could choose to take JD-Next for a preview of law school and then use the exam for admissions purposes. Admitted students may also want to participate for skill development before starting law school. In the second situation, schools could also use the exam to help target students who are already admitted but could benefit from academic support services.

The JD-Next program and exam have been offered through five summer cohorts starting in 2019. The exam's reliability, validity, and fairness in the initial two cohorts are described in prior work (Findley et al., 2023). That published research used data from a nationally recruited cohort of students enriched for racial and ethnic diversity in 2019 and 17 participating law schools in 2017. The researchers provided performance-based incentives to encourage test takers to exert effort on the exam, though it was delivered only for research purposes. In terms of reliability, after excluding five items, the MC portion of the JD-Next exam was found to have a Cronbach's alpha of .85. For the essay question, after training and development and revision of a rubric, two graders achieved an interrater reliability of 95.9%. The researchers

also examined content validity and construct validity and found them to be satisfactory, ensuring that the exam measured what was intended.

In terms of predictive validity, the researchers sought to establish that the exam predicted first-semester law school grade point average (LGPA; Findley et al., 2023). A raw correlation was observed for the 2019 cohort, $r = .480$ ($N = 62$), and for the 2020 cohort, $r = .415$ ($N = 238$). The authors also calculated multivariate regressions, with base models including each law school's median LSAT score (as a measure of selectivity) and the student's undergraduate grade point average (UGPA). In 2019, the addition of the JD-Next to this base model significantly predicted LGPA, $r = .542$, and the test accounted for an additional 16.7% variance in students' first-year LGPAs, $p < .001$. Likewise, in 2020, the addition of the JD-Next exam significantly predicted LGPA, $r = .510$, and accounted for an additional 12.2% variance in students' first-year LGPAs, $p < 0.001$. As a benchmark, the authors compared the JD-Next exam scores with LSAT scores (from both the LSAT exam itself and a few LSAT scores converted from GRE scores, using the ETS conversion tool). The authors replicated known validity estimates for the LSAT in published reports and found that the JD-Next exam had similar predictive power. In some cases, the JD-Next scores provided incremental validity even on top of the power provided by the LSAT scores.

For the larger 2020 cohort, the authors also examined validity and incremental validity in subsets of law schools according to selectivity (top 50, 50–100, and 100+ ranks; Findley et al., 2023). In top-50 schools, the authors found that the JD-Next exam and LSAT both provide statistically significant incremental validity. For schools in the 51–100 and 100+ ranks, the JD-Next score provides a statistically significant improvement in predicting LGPA, but the LSAT's incremental predictive power is smaller and cannot be distinguished from the null.

Findley et al. (2023) also examined score disparities between racial and ethnic groups on both the LSAT and the JD-Next exam. Their data replicated known disparities in the LSAT scores but found that for the JD-Next exam, the disparities were substantially smaller, and nonsignificant in the 2020 cohort.

This study builds on the prior work in several ways. First, to improve the security of the testing environment, the test designers have developed additional test items and implemented a linear-on-the-fly model with pseudo-random item selection so that each test taker saw a different subset of items. Second, we introduced a new scoring and reliability estimation methodology, using item response theory (IRT; Lord & Novick, 1968). Third, we present validity and fairness findings based on the 2021 cohort, serving as a replication of other analysts' prior work on the 2019 and 2020 cohorts (Findley et al., 2023). Fourth, the present study relies on first-year GPAs as the outcome variable, whereas prior work has focused on first-semester GPAs. Fifth, for underrepresented groups (URGs), we not only examine score disparities and validity (as in prior work) but also report on incremental validity, beyond UGPA and law school selectivity for those groups, to understand whether the assessment provides meaningful additional information for these populations. Finally, we examine whether students are the first in their families to attend college or professional school ("first generation"). These analyses will primarily establish if there is sufficient support for using the JD-Next examination for admitted students, the first stated purpose of the examination, but could provide preliminary evidence to support use in law school admissions.

Instrument

The JD-Next version used in this study consisted of a pool of 87 MC items, from which every examinee received 60 items. Every examinee received a common block of six items, which served as anchor items to allow for putting all administered versions of the test on the same scale. The remaining 54 MC items for each examinee were drawn from the 81 unique items left in the pool. Using this method, every examinee received a unique set of items. Additionally, every examinee responded to two essays. These essays were scored analytically using criteria that fell into four categories: issue, rule, application, and conclusion for Essay 1 and issue, element, application, and conclusion for Essay 2. The exam was hosted on the JD-Next course learning management system (LMS; Brightspace, D2L). The exam was open-book, and students could use any resources or materials from the course for reference while taking the exam. Students had 3 hours (180 min) to complete the exam, but additional time accommodations could be given to individual students. Students were advised to spend between 20 and 30 min of the exam time on the essay. They were also instructed to write clearly and to use IRAC style (e.g., a legal writing style that reviews the legal issue, the legal rule, the analysis, and the conclusion, in that order). For the MC section, students had approximately 2.5 min per question. Students were advised in the instructions to select the best

possible answer, that is, one that was most complete and accurate. Students were not able to return to previously answered questions.

Data Collection

An item bank was developed, with MC questions for each exam topic and two essay questions. Fifteen topics on the exam covered skills and doctrinal content, and each topic had between four and eight questions. The LMS software randomly selected four MC questions from each topic and one of the essay prompts as a stratified sample for presentation to each test taker. The LMS software recorded the individual student responses to the 60 MC questions and one essay, and the data were exported for analysis.

Sample Construction

By an e-mail to deans and admissions officers, all 200 accredited American law schools were invited to participate in JD-Next in summer 2021 by signing memoranda of understanding, agreeing to open the course to their incoming students, either only those who were admitted or both those admitted and waitlisted, at the school's discretion. The course was provided free of charge to both schools and students. Students were incentivized to complete the course and perform well on the exam, as in Findley et al. (2023). Incentives included gift cards for department stores and restaurants (e.g., Starbucks and Target) in varying amounts as well as items like textbooks, video game consoles, fitness trackers, and mugs. Students who engaged in the course by completing assignments or participating in course activities (e.g., a midcourse survey or discussion boards) were entered into drawings for the incentive prizes. In total, 26 schools participated in 2021, and 15 schools submitted complete data in time for the present validity analysis. In 2021, none of the participating schools used the JD-Next scores for admissions decisions. The 15 schools were from all regions of the country and varied in selectivity. The number of students from each school ranged from one student to 46 students.

This research project was determined to be exempt by the Human Subjects Protection Program at the University of Arizona. Participating students provided informed consent and waived educational privacy, authorizing their law schools to release first-year grades to the JD-Next research team. Each school registrar submitted data directly to the JD-Next team. For schools that did not use the standard 4-point grading scale, we converted grades using schools' official crosswalk guidance, according to the corresponding schools' websites. The investigators at University of Arizona matched the data and provided deidentified data to ETS for analysis.

For our analyses, the sample ($N = 240$) was limited to the students who took the JD-Next exam, matriculated in law school, and completed their first year of law school (1L year) so as to provide grades as an outcome variable. The sample included students matriculating at law schools ranked by *U.S. News and World Report* (2022) as in the top 100 ($n = 117$, 49%) and schools ranked outside the top 100 ($n = 123$, 51%). For analyses on score disparities, outcomes data are unnecessary, so the full sample was larger ($N = 350$). The sample included 188 women (54%) and 162 men (46%). With respect to race and ethnicity, we highlight larger groups that facilitate analysis, specifically students identifying as Black/African American ($n = 28$, 8%), Hispanic ($n = 59$, 17%), and White non-Hispanic ($n = 241$, 69%). All racial and ethnic groups, except White non-Hispanic and Asian students, are pooled in an URG variable ($n = 100$, 29%). Our sample contained a small group of Asian students ($n = 24$, 6%) who were not included in these analyses. For fairness in analyses looking at predictive validity, only the URG subpopulation is large enough for analysis, with 43 students (18% of the $N = 240$ validity sample) having both test scores and grades.

Scoring and Reliability

For previous cohorts, all students responded to the same set of questions. On the basis of this design, the JD-Next exam was scored using classical scoring methods, for simplicity of score exports from the testing platform. However, for the summer 2021 cohort, the design was modified to collect data for a larger pool of items. Under this new design, all students received a common block of 6 MC items, followed by 54 randomly assigned MC items (from the pool of 87 MC items), such that every examinee received a unique set of items. Furthermore, under the previous scoring model, results were compiled using only the MC items. For the updated exam, students' scores are based on a combination of results from the MC items and two essays. As noted previously, the essays were scored analytically using criteria that fell into four categories: issue,

rule, application, and conclusion for Essay 1 and issue, element, application, and conclusion for Essay 2. On the basis of the scoring rubric, each of these categories included multiple criteria that were scored individually by raters (25 criterion scores). Treatment of these criterion scores in the scoring model is discussed later.

When computing a total score (a sum of the number of correct items) for the MC items, there is an implicit assumption that the difficulty of the unique item sets is the same for all students; however, due to chance, some students could receive a notably harder or easier set of items, resulting in scores that appear to be lower or higher, respectively. To account for any potential differences, an IRT (Lord & Novick, 1968) model can be used to create a set of calibrated item parameters, and scores, that are all on the same scale. For this assessment, the MC items were calibrated using the two-parameter logistic model (2PL; Birnbaum, 1968). How, then, are the essay scores handled? One approach would be to create a separate essay score as a supplement to the MC scores; however, given that the essays are part of the intended construct, a decision was made to incorporate the criterion scores into the estimates of student performance. An iterative approach was used to evaluate the criterion scores within an IRT framework, treating the criterion scores as independent dichotomous items and recoding the items into various polytomous subsets (due to dependencies among the criterion scores). Ultimately, polytomous codings were used for each of the four categories, for each essay, identified previously. These items were calibrated concurrently with the MC items using the generalized partial credit model (GPCM; Muraki, 1992) such that all the MC items, recoded essay scores, and overall student scores are on the same scale. Item parameters for the 2PL and GPCM were estimated using the software MDLTM (von Davier, 2016).

After the initial scaling with the MC items and essay scores was complete, the results were evaluated to ensure estimation convergence and to identify potentially problematic items. With respect to the latter, we used the following criteria to flag problematic items: (a) items with (unscaled) difficulty values greater than 5.0 or less than −5.0 and/or items with negative slopes or slopes less than 0.1 and (b) items with root-mean-square difference item fit statistics greater than 0.15. Four items were flagged and excluded due to negative slopes. The item parameters for the remaining items were all within acceptable ranges and fit the data well.

The end result of this calibration was the creation of a unidimensional scale with expected a posteriori student abilities estimated based on the final item parameters. To be clear, the item parameter estimation was implemented in two different ways: first without the essay questions included, then with the essay questions included. When the essay questions were not included, the estimated marginal reliability was .82; this is considered adequate (Nunnally, 1978). When the essay questions were included, the estimated marginal reliability with the combined MC items and essay scores was .87.

Validity Methodology

To explore the validity of the JD-Next assessment, in the next two sections, the methodology applied in Findley et al. (2023) is used and described. For all models, we sought to understand not only whether the exam provided a valid prediction of LGPA but also whether it provided *incremental* predictive value above and beyond the other information that would be available to a law school admissions office, which would include UGPA and LSAT score. Although students were nested within schools, multilevel modeling was not utilized due to the small number of schools in this study. As an alternative, the median LSAT scores for participants' schools were included in the regression models to account for the differing selectivities and academic environments of the participants' law schools. For all analyses, the LSAT scores used include a mixture of LSAT scores and GRE scores converted to LSAT scores based on Klieger et al. (2018) and are referred to as "LSAT (including converted GRE)."

As described in the previous section, two JD-Next scores were derived, one based on the complete JD-Next exam, which included MC items and an essay question, and one based solely on the MC items. To understand whether the essay question provided any additional incremental validity, initial predictive validity models (labeled as Model 1) using the two sets of scores explored the incremental predictive power of each in predicting LGPA, beyond the use of only UGPA and median LSAT as predictors. Results of the models are shown in Table 1.

On the basis of these analyses, the point estimates for JD-Next MC-only scores suggest slightly greater validity and incremental validity as compared to overall scores, which include the essay scores. When considering this and the finding that the reliabilities of both scores were adequate, in addition to the complexities of administering and scoring the essay, using the MC items alone seems to be sufficient. Given this, for remaining sections of this report, the JD-Next score will refer to the score based only on the MC items. However, given the initial intention to include the essay in the JD-Next

Table 1 Comparison of Incremental Validity Using JD-Next Overall and Multiple-Choice Scores in Predicting First-Year Law School Grade Point Average

	Overall (MC + essay)			MC only		
Variable(s) entered	R	R^2	ΔR^2	R	R^2	ΔR^2
Model 1						
UGPA and mLSAT	.35	.12		.35	.12	
JD-Next	.49	.24	.12*	.53	.28	.16*

Note. MC = multiple-choice. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. *$p < .01$.

Table 2 Correlation Matrix of Variables Included in the Hierarchical Regression Analysis of Students' First-Year Law School Grade Point Averages

	mLSAT	LSAT	JD-Next	LGPA
UGPA	.34*	.06	.03	.17*
mLSAT		.61*	.30*	.34*
LSAT			.44*	.52*
JD-Next MC				.48*

Note. $N = 240$. LGPA = law school grade point average. LSAT = Law School Admission Test. MC = multiple-choice. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. *$p < .01$.

score and other potential reasons to include the essays, for example, construct representation or institutions' interest in the essays, analyses using the overall score are presented in the appendix.

Using the MC-only version of JD-Next, we fit several models to better understand the relationships between predictor and outcome variables. As a first step, we examined the correlations of all the variables included in the models. After exploring the correlations, four models were fit for the complete sample, and all subsamples were examined. Model 1 was run as described in the previous analysis. As a base for comparison, Model 2 included the traditional measure, the LSAT score, in predicting LGPA, in place of the JD-Next score, also looking at the incremental validity beyond the use of only UGPA and median LSAT. Alternatively, the JD-Next score may be useful as a complement to other standardized test scores, allowing greater predictive power even if a student already has another measure, such as LSAT or GRE score (and vice versa). In Model 3, the JD-Next exam score was entered, followed by the LSAT (including converted GRE) score, to determine whether using both scores improved prediction. The order of entry for predictor variables was reversed in Model 4, with the LSAT (including converted GRE) score entered first, followed by the JD-Next exam score. Models 3 and 4 address the question of whether the JD-Next exam score can supplement traditional law school entrance examination scores.

Validity Findings

Correlations for all variables used in the regression models are shown in Table 2. There are moderate relationships among many of the variables. Not surprisingly, JD-Next scores were moderately correlated with LSAT (including converted GRE) scores, though low enough to warrant further exploration to show if each individually contributes to the prediction of first-year LGPA.

Taking account of other variables, following the methodology discussed earlier, we used a series of linear regression analyses with a predetermined order of variable entry, as shown in Table 3.

Model 1 shows that adding the JD-Next exam score to the law school median LSAT (including converted GRE) score and UGPA significantly increases the variance explained in first-year GPA, accounting for an additional 16% variance in students' first-year LGPA. Model 2 shows that students' LSAT (including converted GRE) scores also increased the variance explained in first-year LGPA, accounting for an additional 17% of the variance in LGPA. When using both the JD-Next score and LSAT score to predict LGPA, order of entry into the model did not impact the results. Whereas Model 3 shows the JD-Next score being entered into the model first, followed by the LSAT (including converted GRE) score, and

Table 3 Summary of Hierarchical Regression Analysis for the JD-Next Exam or Law School Admission Test (Including Converted GRE) With Undergraduate Grade Point Average and Median Law School Admission Test (Including Converted GRE) Score Predicting First-Year Law School Grade Point Average

Variable(s) entered	R	R^2	ΔR^2
Model 1			
UGPA and mLSAT	.35	.12	
JD-Next MC	.53	.28	.16*
Model 2			
UGPA and mLSAT	.35	.12	
LSAT	.54	.29	.17*
Model 3			
UGPA and mLSAT	.35	.12	
JD-Next MC	.53	.28	.16*
LSAT	.61	.37	.09*
Model 4			
UGPA and mLSAT	.35	.12	
LSAT	.54	.29	.17*
JD-Next MC	.61	.37	.08*

Note. $N = 240$. LSAT = Law School Admission Test. MC = multiple-choice. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. $^*p < .01$.

Model 4 shows the JD-Next exam entered after the LSAT (including converted GRE), both additions led to statistically significant increases in predictive power after the other predictor was added.

Our analyses are broadly consistent with past research supporting the JD-Next exam as a valid predictor of law school performance, to a similar degree compared to the standardized admissions tests. Additionally, our findings suggest that the JD-Next score may also be useful as a supplemental admissions tool even for students who have taken one of the standardized admissions tests (the LSAT or the GRE).

Validity by School Selectivity

Among the 199 American Bar Association (ABA)-accredited law schools in the United States are a wide range of academic profiles. For the 2021 cohort, which was recruited from a diverse group of 15 ABA-accredited law schools, we sought to understand the validity of the JD-Next exam across that range. To allow for statistical power and to avoid identifying any particular school, we grouped the participating schools into two roughly equally sized categories: those ranked in the top 100 (median LSAT scores of 154–180; Group I) and those outside the top 100 (median LSAT scores of 144–153; Group II). Median LSAT score was used as a measure of school selectivity. For cases in which a student transferred from one law school (where they participated in JD-Next) to another (which may or may not have participated in JD-Next), we counted that student in the group in which they matriculated, if we could secure their first-year grades.

For both groups, as earlier, we examined the validity of both the JD-Next exam and the LSAT (including converted GRE). We found a positive correlation between both tests' scores and law school grades in both of the school groups. More importantly, we tested for incremental validity of these variables, above that provided by median LSAT and UGPA. As Table 4 shows, when examining the ΔR^2 values for both groups, both the JD-Next exam and the LSAT provide statistically significant improvements in predicting LGPA in all the models.

However, for Group I schools, as shown in Table 4, the base model of UGPA and median LSAT appears to have more predictive power than for Group II schools (although no statistical significance tests were run). Nonetheless, the JD-Next score provided a statistically significant improvement in predicting LGPA in all the models. Even though the LSAT's incremental predictive power in Model 3 appears smaller for Group II, the JD-Next seems to provide greater incremental validity for this group. (Again, no significance tests were performed.)

Score Disparities for Racial and Ethnic Groups

Although our validity models do not use race or ethnicity as a covariate, we were interested in the performance on the JD-Next exam of students from diverse populations. In particular, it was important to explore if students from historically

Table 4 Summary of Hierarchical Regression Analysis for the JD-Next Exam or Law School Admission Test (Including Converted GRE) With Undergraduate Grade Point Average and Median Law School Admission Test (Including Converted GRE) Score Predicting First-Year Law School Grade Point Average by School Grouping

	Group I: Top 100[a]			Group II: Outside the top 100[b]		
Variable(s) entered	R	R^2	ΔR^2	R	R^2	ΔR^2
Model 1						
UGPA and mLSAT	.46	.21		.33	.11	
JD-Next MC	.61	.37	.16*	.53	.28	.17*
Model 2						
UGPA and mLSAT	.46	.21		.33	.11	
LSAT	.66	.43	.22*	.47	.22	.12*
Model 3						
UGPA and mLSAT	.46	.21		.33	.11	
JD-Next MC	.61	.37	.16*	.53	.28	.17*
LSAT	.69	.48	.11*	.57	.33	.06*
Model 4						
UGPA and mLSAT	.46	.21		.33	.11	
LSAT	.66	.43	.22*	.47	.22	.12*
JD-Next MC	.69	.48	.05*	.57	.33	.11*

Note. LSAT = Law School Admission Test. MC = multiple-choice. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. [a] $n = 117$. [b] $n = 123$. *$p < .01$.

underrepresented populations, especially larger groups, such as Black/African American and Hispanic, tend to score lower on the JD-Next exam and if the exam's validity is robust across these groups and other, smaller groups. These questions about score disparities are important because use of admissions tools can impact efforts to increase diversity in law schools. If admissions officers use a particular tool in deciding which applicants to admit, and examinees from one or more racial or ethnic groups score lower on average on that tool compared to their peers from other racial and ethnic groups, then members of those lower-scoring groups are more likely to be denied admission, and their overall representation in law school and potentially in the legal profession could thereby be reduced. Aside from simple score disparities, we were also interested in examining the predictive validity for students across racial and ethnic groups.

For this purpose, we compared subgroups' performance on the JD-Next exam to their performance on the LSAT (including GRE scores converted to LSAT scores). When comparing scores, we observed differences in the scores depending on race/ethnicity for both tests. Notably, our data replicate some of the same score disparities shown by the LSAC for the LSAT exam from previous years (Dalessandro et al., 2014; Lauth & Sweeney, 2022). For example, in our study, White (non-Hispanic) test takers ($n = 240$) scored 158.46 ($SD = 6.64$) on the LSAT on average, whereas Black/African American test takers ($n = 28$) scored 149.54 ($SD = 5.94$) on average. This difference of about 9 points is very similar to the 10-point and 11-point differences reported in various years by the LSAC, based on its comprehensive census of test taker data. Hispanic test takers ($n = 59$) scored 156.53 ($SD = 4.85$) on average. The general pooled URG ($n = 100$), including all races and ethnicities, except White non-Hispanic and Asian students (i.e., Hispanic, Black/African American, Native American and Alaska Native, multiracial), scored 154.18 ($SD = 6.19$), achieving lower scores than White (non-Hispanic) test takers. Similar differences for the JD-Next exam are harder to interpret; however, similar patterns are seen.

Given that the LSAT (including converted GRE) and the JD-Next have different scales and different score distributions, it is necessary to use standardized statistics to evaluate the impact of these differences and to compare them across groups and exams. Cohen's *d* statistic was used here so that the differences were described in the same units irrespective of the score scale, following an approach by Camara and Schmidt (1999).

Using Cohen's *d*, for Black/African American, Hispanic, and URG test takers compared with White test takers, we found sizable disparities in LSAT (including converted GRE) test scores, and in every case, the point estimates for the JD-Next exams trend toward smaller disparities. On the other hand, gender disparities were slightly larger for the JD-Next exam than for the LSAT (including converted GRE), but this difference was relatively smaller. Table 5 displays Cohen's *d* statistics for the two tests, with sample sizes and 95% confidence intervals that account for unavoidable uncertainty in estimating effect sizes.

Table 5 Standardized Score Disparities for Underrepresented Groups on the JD-Next Exam and Law School Admission Test

	Focal group			Reference group					95% CI for Cohen's *d*	
Group (focal vs. reference)	*n*	*M*	*SD*	*n*	*M*	*SD*	*M* (focal) − *M* (reference)	Cohen's *d*	Lower	Upper
Female vs. male										
LSAT/GRE	188	156.18	6.61	162	157.96	6.87	−1.78	−0.26	−0.48	−0.05
JD-Next MC	189	0.54	0.36	162	0.65	0.34	−0.11	−0.31	−0.52	−0.10
Black/African American vs. White (non-Hispanic)										
LSAT/GRE	28	149.54	5.94	240	158.46	6.64	−8.92	−1.36	−1.62	−1.09
JD-Next MC	28	0.52	0.33	241	0.63	0.36	−0.11	−0.30	−0.54	−0.06
Hispanic vs. White (non-Hispanic)										
LSAT/GRE	59	156.53	4.85	240	158.46	6.64	−1.93	−0.31	−0.53	−0.08
JD-Next MC	59	0.54	0.32	241	0.63	0.36	−0.09	−0.25	−0.47	−0.02
URG vs. White (non-Hispanic)										
LSAT/GRE	100	154.18	6.19	240	158.46	6.64	−4.28	−0.66	−0.88	−0.44
JD-Next MC	100	0.54	0.32	241	0.63	0.36	−0.09	−0.25	−0.46	−0.04

Note. LSAT = Law School Admission Test. MC = multiple-choice. URG = underrepresented group. The LSAT variable includes converted GRE scores. URG includes Hispanic, Black/African American, Native American and Alaska Native, and multiracial.

Table 6 Predictive Validity for Underrepresented Students and Other Students

	URG[a]			Other students[b]		
Variable(s) entered	*R*	R^2	ΔR^2	*R*	R^2	ΔR^2
Model 1						
UGPA and mLSAT	.26	.07		.37	.13	
JD-Next MC	.47	.23	.16**	.55	.30	.17**
Model 2						
UGPA and mLSAT	.26	.07		.37	.13	
LSAT	.59	.35	.28**	.50	.25	.11**
Model 3						
UGPA and mLSAT	.26	.07		.37	.13	
JD-Next MC	.47	.23	.16**	.55	.30	.17**
LSAT	.64	.42	.19**	.59	.35	.05**
Model 4						
UGPA and mLSAT	.26	.07		.37	.13	
LSAT	.59	.35	.28**	.50	.25	.11**
JD-Next MC	.64	.42	.07*	.59	.35	.10**

Note. LSAT = Law School Admission Test. MC = multiple-choice. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. URG = underrepresented group. The LSAT variable includes converted GRE scores. [a] $n = 43$. [b] $n = 197$. *$p < .05$. **$p < .01$.

We also examined the predictive validity of both tests for URG test takers separately. Table 6 displays R^2 values for models including UGPA, LSAT (including converted GRE), and school median LSAT (including converted GRE), grouping students into two groups: URG, as described earlier, and other students, defined as students not in the URG categorization (White non-Hispanic and Asian students).

The JD-Next exam demonstrates consistent incremental predictive validity across both groups over other measures, as shown by similar incremental R^2 values for Model 4. Additionally, a model adding a URG indicator variable, using the complete sample, was examined to investigate potential differential prediction (i.e., predictive bias, differences in slopes and intercepts; Cleary, 1968).[1] This model showed that the interaction term between the URG indicator and JD-Next was not significant, $p = .33$, meaning there was no significant difference in slopes. Figure 1 depicts the minimal difference in slopes between the two groups. Additionally, an exploration of the differences between intercepts showed a difference in

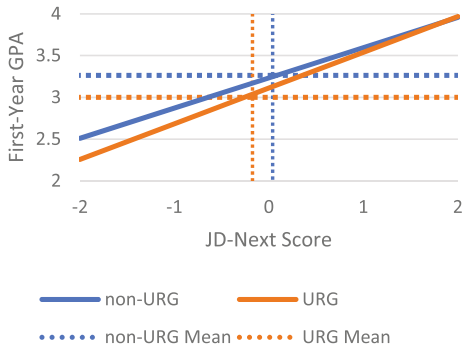

Figure 1 Relationships between JD-Next score and first-year grade point average by underrepresented group status. URG = underrepresented group.

Table 7 Standardized Score Disparities for First-Generation Student Status on the JD-Next Exam and Law School Admission Test

	Focal group			Reference group					95% CI for Cohen's *d*	
Group (focal vs. reference)	*n*	*M*	*SD*	*n*	*M*	*SD*	*M* (focal) − *M* (reference)	Cohen's *d*	Lower	Upper
First generation vs. parent with bachelor's degree										
LSAT/GRE	65	155.78	6.09	191	158.70	6.96	−2.92	−0.43	−0.68	−0.18
JD-Next MC	65	0.51	0.39	192	0.65	0.35	−0.14	−0.41	−0.65	−0.16

Note. LSAT = Law School Admission Test. MC = multiple-choice. The LSAT variable includes converted GRE scores.

intercepts between the two groups approaching significance, $p = .06$. On the basis of this finding, further investigation showed that this difference is based on a significant difference in means between the two groups in LGPA, $p < .01$, and not JD-Next, $p = .19$, which can be considered evidence of potential criterion bias, not test bias, and thus does not preclude using the examination (Meade & Fetzer, 2009). The differences in the means for the two measures for the two groups are also shown in Figure 1.

In summary, the JD-Next exam showed consistently smaller score disparities for underrepresented students compared to the LSAT. We found positive predictive validity for both exams for underrepresented test takers.

Score Disparities Based on Parental Education

We were also interested in the performance on the JD-Next exam of students from households in which they were the first to complete a bachelor's degree (first generation).[2] In particular, it was important to explore if these students tend to score lower on the JD-Next exam and if the exam's validity is robust across this group.

For this purpose, we compared subgroup differences in performance on the JD-Next exam and performance on the LSAT (including GRE scores converted to LSAT scores). We observed similar differences in the scores depending on first-generation status for both tests, using the Cohen's *d* statistic. Table 7 displays Cohen's *d* statistics for the two tests, with sample sizes and 95% confidence intervals that account for unavoidable uncertainty in estimating effect sizes.

We also examined the predictive validity of both tests for first-generation test takers separately. Table 8 displays R^2 values for models including UGPA, LSAT (including converted GRE), and school median LSAT (including converted GRE).

The JD-Next exam demonstrates consistent incremental predictive validity across both groups over other measures, as shown by similar incremental R^2 values for Model 4. Additionally, a model using the complete sample, adding a first-generation status indicator variable, was fit to investigate potential differential prediction (i.e., predictive bias, differences in slopes and intercepts; Cleary, 1968). This model showed that the interaction term between the first-generation status indicator and JD-Next was not significant, $p = .57$, meaning there was no significant difference in slopes. Additionally, an exploration of the differences between intercepts showed no difference in intercepts between the two groups, $p = .45$. On the basis of this finding, no differential prediction was found.

Table 8 Predictive Validity for First-Generation Students and Other Students

	First generation[a]			Other students[b]		
Variable(s) entered	R	R^2	ΔR^2	R	R^2	ΔR^2
Model 1						
UGPA and mLSAT	.32	.10		.35	.12	
JD-Next MC	.54	.29	.19**	.51	.26	.13**
Model 2						
UGPA and mLSAT	.32	.10		.35	.12	
LSAT	.60	.37	.26**	.51	.26	.13**
Model 3						
UGPA and mLSAT	.32	.10		.35	.12	
JD-Next MC	.54	.29	.19**	.51	.26	.13**
LSAT	.64	.43	.14**	.57	.33	.07**
Model 4						
UGPA and mLSAT	.32	.10		.35	.12	
LSAT	.60	.37	.26**	.51	.26	.13**
JD-Next MC	.66	.43	.07*	.57	.33	.07**

Note. LSAT = Law School Admission Test. MC = multiple-choice. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. [a] $n = 60$. [b] $n = 178$. *$p < .05$. **$p < .01$.

In summary, both the JD-Next exam and the LSAT showed score disparities for first-generation students. We found positive predictive validity for both exams for first-generation test takers.

Conclusions and Discussion

Law schools are in need of reliable, valid, and unbiased tools to help predict the success of applicants in completing their JD degree programs. Although traditional standardized assessments like the LSAT and GRE are widely used, alternative approaches focusing on discipline-specific content and providing a supportive learning environment address limitations of general tests assessing various abilities. The JD-Next program and exam meet these needs. Building on prior work by introducing new scoring and reliability estimation methodology, on the basis of our analyses, this study provides evidence that the JD-Next exam is reliable and valid for predicting law school success for admitted students, yielding a significant increment in predictive power over UGPA and LSAT (including converted GRE) scores, while controlling for school median LSAT (including converted GRE) scores. Additionally, we found it to show smaller score disparities than other measures and to be equally predictive for students from underrepresented minority groups, demonstrating that it is a fair measure for students from all backgrounds.

Ongoing review of test design to ensure reliability, validity, and fairness of the assessment is recommended, particularly if there is a desire to continue administering the essay question. Nonetheless, these findings, in conjunction with previous findings, support the use of the JD-Next exam, particularly for supporting admitted students' readiness for law school by teaching critical skills in case briefing and legal analysis ahead of law school.

However, it is worth noting that the JD-Next program and exam have also been proposed for use in law school admissions. While the current analyses and collected data cannot directly support this particular use at present, implementing range restriction adjustments could provide insights into the potential applicability of these results to the law school applicant population (Thorndike, 1949). Because this population is likely to have a greater variance in scores on both the LSAT and the JD-Next exam, validity coefficients would likely be larger, providing stronger validity evidence for this purpose. Additionally, it is important to consider other potential implications of utilizing the JD-Next exam for admissions, as this would be the first instance of employing the exam in a high-stakes scenario. Thus further exploration and data collection under these circumstances are encouraged to gather more robust validation evidence. Meanwhile, the exam could be incorporated as one measure with a minimal weight along with multiple other measures as part of the admissions process. This approach would provide an opportunity to explore the use of this program and exam for admissions while collecting additional data for further validation efforts.

Furthermore, supplemental research can be conducted using additional advanced methodologies or data to address the limitations encountered in this study. Future data collection should ideally include high-stakes administration of the

JD-Next exam to law school applicants. Moreover, it is important to note that data were not available for admitted students who took the JD-Next but may have dropped out or not completed their first year of law school. Imputation could be used to estimate outcome data for these students. Similarly, alternative models could be applied, embedding background variables like race/ethnicity and first-generation status into the initial models, to explore the significance of the grouping variables being explored. It would also be advantageous to complete a research study, potentially using a propensity score matching design, to establish any benefits of students being exposed to the JD-Next program and exam. Finally, additional outcome data could provide insights into the potential long-term effects of the JD-Next program and exam. These future research endeavors will help contribute to the growing body of evidence on the efficacity of the JD-Next program and exam for both potential and admitted law school students.

Acknowledgments

The authors gratefully acknowledge ETS for support of the 2021 and 2022 cohorts of the JD-Next program, including support of this research. In addition, AccessLex Institute is gratefully acknowledged for other financial support of the program.

Notes

1 Because the URG is relatively small ($n = 43$), power calculations were performed to confirm that this methodology was appropriate. These calculations revealed that the sample sizes in our study were sufficient for analyses to proceed (Shieh, 2018).
2 It should be noted that through chi-square analyses, we detected a slight overlap between first-generation status and URG membership, small enough to warrant separate analyses using these two variables.

References

Amabebe, E. M. (2020). Beyond "valid and reliable": The LSAT, ABA Standard 503, and the future of law school admissions. *NYU Law Review*, *95*(6), Article 1860.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

Buzick, H. M., Robertson, C., Findley, J. D., Burross, H. L., Charles, M., &; Klieger, D. M. (2023). The association of participation in a summer prelaw training program and first-year law school sudents' grades. *Journal of Educational Research and Practice*, *13*(1), 181–202. https://doi.org/10.5590/JERAP.2023.13.1.14

Camara, W. J., & Schmidt, A. E. (1999). *Group differences in standardized testing and social stratification* (Report No. 99-5). College Entrance Examination Board.

Cheng, K. C., Findley, J., Cimetta, A., Burross, H. L., Charles, M., Balser, C., Li, R., & Robertson, C. T. (in press). JD-Next: A randomized experiment of an online scalable program to prepare diverse students for law school. *Journal of Legal Education*.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*(2), 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Dalessandro, S. P., Anthony, L. C., & Reese, L. M. (2014). *LSAT performance with regional, gender, and racial/ethnic breakdowns: 2007–2008 through 2013–2014 testing years* (LSAT Technical Report No. 14-02). Law School Admissions Council.

Den Hartigh, R. J. R., Niessen, S. M., Frencken, W. G., & Meijer, R. R. (2018). Selection procedures in sports: Improving predictions of athletes' future performance. *European Journal of Sport Science*, *18*(9), 1191–1198. https://doi.org/10.1080/17461391.2018.1480662

Ever-Tite Roofing Corp. v. Green, 83 So.2d 449 (1955).

Farley, A. N., Swoboda, C. M., Chanvisanuruk, J., McKinley, K. M., Boards, A., & Gilday, C. (2019). A deeper look at bar success: The relationship between law student success, academic performance, and student characteristics. *Journal of Empirical Legal Studies*, *16*(3), 605–629. https://doi.org/10.1111/jels.12228

Findley, J., Cimetta, A., Burross, H. L., Cheng, K. C., Charles, M., Balser, C., Li, R., & Robertson, C. (2023). JD-Next: A valid and reliable tool to predict diverse students' success in law school. *Journal of Empirical Legal Studies*, *20*(1), 134–165. https://doi.org/10.1111/jels.12342

Guinier, L. (2015). *The tyranny of the meritocracy: Democratizing higher education in America*. Beacon Press.

Hamer v. Sidway, 124 N.Y. 538, 27 N.E. 256 (N.Y. 1891).

Kidder, W. C. (2001). Does the LSAT mirror or magnify racial and ethnic differences in educational attainment? A study of equally achieving "elite" college students. *California Law Review*, *89*(4), 1055–1124. https://doi.org/10.2307/3481291

Klieger, D. M., Bridgeman, B., Tannenbaum, R. J., & Cline, F. A. (2016). The validity of GRE® scores for predicting academic performance at the University of Arizona James E. *Rogers College of Law.* https://online.wsj.com/public/resources/documents/gre_validitystudy_arizona.pdf

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). ETS. https://doi.org/10.1002/ets2.12213

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, *315*(5815), 1080–1081. https://doi.org/10.1126/science.1136618

Lauth, L. A., & Sweeney, A. T. (2022). *LSAT performance with regional, gender, racial, and ethnic breakdowns 2011–2012 through 2017–2018 testing years* (Technical Report No. TR 22-01). Law School Admissions Council.

Lidz, C. (1995). Dynamic assessment and the legacy of L. S. Vygotsky. *School Psychology International*, *16*(2), 143–153. https://doi.org/10.1177/0143034395162005

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Meade, A. W., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods*, *12*(4), 738–761. https://doi.org/10.1177/1094428109331487

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Predicting performance in higher education using proximal predictors. *PLoS One*, *11*(4), Article e0153663. https://doi.org/10.1371/journal.pone.0153663

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLoS ONE*, *13*(6), Article e0198746. https://doi.org/10.1371/journal.pone.0198746

Nunnally, J. C. (1978). An overview of psychological measurement. In B. B. Wolman (Ed.), *Clinical diagnosis of mental disorders* (pp. 97–146). Springer. https://doi.org/10.1007/978-1-4684-2490-4_4

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, *58*(4), 1009–1037. https://doi.org/10.1111/j.1744-6570.2005.00714.x

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Shieh, G. (2018). Power and sample size calculations for comparison of two regression lines with heterogeneous variances. *PLoS ONE*, *13*(12), Article e0207745. https://doi.org/10.1371/journal.pone.0207745

Shultz, M. M., & Zedeck, S. (2012). Admission to law school: New measures. *Educational Psychologist*, *47*(1), 51–65. https://doi.org/10.1080/00461520.2011.610679

Sternberg, R. J., & Grigorenko, E. L. (2020). *Dynamic testing: The nature and measurement of learning potential*. Cambridge University Press.

Taylor, A. N. (2014). Reimaging merit as achievement. *New Mexico Law Review*, *44*(1), 1–47.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. John Wiley.

Tucker, G. C. (2016, October 26). *Putting pre-employment tryouts to the test*. Society for Human Resource Management. https://www.shrm.org/topics-tools/news/hr-magazine/putting-pre-employment-tryouts-to-test

U.S. News and World Report. (2022). *2022–2023 best law schools*. https://www.usnews.com/best-graduate-schools/top-law-schools/law-rankings

von Davier, M. (2016). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. ETS.

Appendix

Validity Analyses Using Overall JD-Next Score (Essay and MC)

As a corollary to Table 2, Table A1 gives the correlations between the overall JD-Next exam and other variables used in the regression models. Correlations with other variables are similar but slightly lower than for the JD-Next MC score.

Table A2 gives the series of linear regression analyses of Table 3, but using the overall JD-Next score.

Model 1 shows that adding the JD-Next exam to the law school median LSAT (including converted GRE) score significantly predicted LGPA and accounted for an additional 12% variance in law students' first-year GPAs, $p < .01$. Model 2 shows that students' LSAT (including converted GRE) scores also significantly predicted LGPA and accounted for an additional 17% of the variance in LGPA. When using both the JD-Next score and LSAT score to predict LGPA, order of entry into the model did not impact the results. Whereas Model 3 shows the JD-Next score being entered into the model first, followed by the LSAT (including converted GRE) score, and Model 4 shows that the JD-Next exam was entered after the LSAT (including converted GRE), both additions were statistically significant.

Table A1 Correlation of Overall JD-Next Score With Variables Included in the Hierarchical Regression Analysis of Students' First-Year 2022 Law School Grade Point Averages

	JD-Next
UGPA	.01
mLSAT	.22*
LSAT	.35*

Note. $N = 240$. LSAT = Law School Admission Test. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. $^{*}p < .01$.

Table A2 Summary of Hierarchical Regression Analysis for the Overall JD-Next Score or Law School Admission Test (Including Converted GRE) Score With Undergraduate Grade Point Average and Median Law School Admission Test (Including Converted GRE) Score Predicting First-Year Law School Grade Point Average

Variable(s) entered	R	R^2	ΔR^2
Model 1			
UGPA and mLSAT	.35	.12	
JD-Next	.49	.24	.12*
Model 2			
UGPA and mLSAT	.35	.12	
LSAT	.54	.29	.17*
Model 3			
UGPA and mLSAT	.35	.12	
JD-Next	.49	.24	.12*
LSAT	.60	.35	.11*
Model 4			
UGPA and mLSAT	.35	.12	
LSAT	.54	.29	.17*
JD-Next	.60	.35	.06*

Note. $N = 240$. LSAT = Law School Admission Test. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. $^{*}p < .01$.

Our analyses here are consistent with the analyses using the JD-Next MC score, with the incremental validity of the overall score being slightly smaller.

Validity by School Groupings

As in Table 4, Table A3 gives the incremental validities for the two groups of schools. When examining the ΔR^2 p-values for both groups, both the overall JD-Next score and the LSAT score provide statistically significant improvements in predicting LGPA in all the models. All findings show similar patterns to the findings for the JD-Next MC scores, though incremental validities are consistently smaller.

Score Disparities for Racial and Ethnic Groups

Similar to analyses for the JD-Next MC score, we compared subgroup differences in JD-Next overall scores to LSAT scores (including GRE scores converted to LSAT scores). Using Cohen's d, for Black/African American, Hispanic, and URG test takers, similar to the JD-Next MC score, in every case, the point estimates for the JD-Next MC score trend toward smaller disparities than the LSAT (including converted GRE). When comparing the disparities here with those of the JD-Next MC score, the disparities are slightly larger for gender and White versus Hispanic but smaller for White versus Black/African American and White versus URG. It should be noted that the White versus Black/African American difference is considerably smaller when using the overall JD-Next, and for this comparison, the scores for the two groups are not significantly different. Table A4 displays Cohen's d statistics for the JD-Next overall score, along with the JD-Next MC score, with sample sizes and 95% confidence intervals that account for unavoidable uncertainty in estimating effect sizes.

Table A3 Summary of Hierarchical Regression Analysis for the Overall JD-Next Score or Law School Admission Test (Including Converted GRE) Score With Undergraduate Grade Point Average and Median Law School Admission Test (Including Converted GRE) Score Predicting First-Year Law School Grade Point Average by School Grouping

	Group I: Top 100[a]			Group II: Outside the top 100[b]		
Variable(s) entered	R	R^2	ΔR^2	R	R^2	ΔR^2
Model 1						
UGPA and mLSAT	.46	.21		.33	.11	
JD-Next	.57	.33	.11*	.49	.24	.14*
Model 2						
UGPA and mLSAT	.46	.21		.33	.11	
LSAT	.66	.43	.22*	.47	.22	.12*
Model 3						
UGPA and mLSAT	.46	.21		.33	.11	
JD-Next	.57	.33	.11*	.49	.24	.14*
LSAT	.68	.47	.14*	.56	.31	.07*
Model 4						
UGPA and mLSAT	.46	.21		.33	.11	
LSAT	.66	.43	.22*	.47	.22	.12*
JD-Next	.68	.47	.04*	.56	.31	.09*

Note. LSAT = Law School Admission Test. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. [a] $n = 117$. [b] $n = 123$. *$p < .01$.

Table A4 Standardized Score Disparities for Underrepresented Groups on JD-Next Overall and Multiple-Choice Scores

	Focal group			Reference group					95% CI for Cohen's d	
Group (focal vs. reference)	n	M	SD	n	M	SD	M (focal) − M (reference)	Cohen's d	Lower	Upper
Female vs. male										
JD-Next overall	189	0.46	0.34	162	0.56	0.36	−0.10	−0.29	−0.50	−0.08
JD-Next MC only	189	0.54	0.36	162	0.65	0.34	−0.11	−0.31	−0.52	−0.10
African American vs. White non-Hispanic										
JD-Next overall	28	0.49	0.36	241	0.54	0.36	−0.05	−0.14	−0.38	0.10
JD-Next MC only	28	0.52	0.33	241	0.63	0.36	−0.11	−0.30	−0.54	−0.06
Hispanic vs. White non-Hispanic										
JD-Next overall	59	0.44	0.33	241	0.54	0.36	−0.10	−0.29	−0.52	−0.06
JD-Next MC only	59	0.54	0.32	241	0.63	0.36	−0.09	−0.25	−0.47	−0.02
URG vs. White non-Hispanic										
JD-Next overall	100	0.47	0.33	241	0.54	0.36	−0.07	−0.21	−0.42	0.00
JD-Next MC only	100	0.54	0.32	241	0.63	0.36	−0.09	−0.25	−0.46	−0.04

Note. MC = multiple-choice. URG = underrepresented group. Pooled URG includes all races and ethnicities showing significant differences from White non-Hispanic students on at least one of the two exams (i.e., Hispanic, Black/African American, Native American and Alaska Native, multiracial).

We also examined the predictive validity of both tests for the URG test takers separately, using the JD-Next overall score. As a corollary to Table 6, Table A5 displays R^2 values for models including UGPA, LSAT (including converted GRE), and school median LSAT (including converted GRE), grouping students by URG.

When the overall JD-Next score is used, the JD-Next exam still has consistent incremental predictive validity across both groups. Additionally, as with the JD-Next MC score, a model adding an URG indicator variable using the complete sample was fit to investigate potential differential prediction (i.e., predictive bias, differences in slopes and intercepts; Cleary, 1968). This model showed that the interaction term between the URG indicator and JD-Next was not significant, $p = .21$, meaning there was no significant difference in slopes. Additionally, an exploration of the differences between

Table A5 Predictive Validity for Underrepresented Students and Others

	URG[a]			Other students[b]		
Variable(s) entered	R	R^2	ΔR^2	R	R^2	ΔR^2
Model 1						
UGPA and mLSAT	.26	.07		.37	.13	
JD-Next	.47	.22	.15**	.51	.26	.13**
Model 2						
UGPA and mLSAT	.26	.07		.37	.13	
LSAT	.59	.35	.28**	.50	.25	.11**
Model 3						
UGPA and mLSAT	.26	.07		.37	.13	
JD-Next	.47	.22	.15**	.51	.26	.13**
LSAT	.66	.43	.21**	.57	.32	.06**
Model 4						
UGPA and mLSAT	.26	.07		.37	.13	
LSAT	.59	.35	.28**	.50	.25	.11**
JD-Next	.66	.43	.08*	.57	.32	.07**

Note. LSAT = Law School Admission Test. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. URG = underrepresented group. The LSAT variable includes converted GRE scores. [a] $n = 43$. [b] $n = 197$. *$p < .05$. **$p < .01$.

Table A6 Standardized Score Disparities for First-Generation Student Status on JD-Next Overall and Multiple-Choice Scores

	Focal group			Reference group					95% CI for Cohen's d	
Group (focal vs. reference)	n	M	SD	n	M	SD	M (focal) − M (reference)	Cohen's d	Lower	Upper
First generation vs. parent with bachelor's degree										
JD-Next overall	65	0.43	0.38	192	0.56	0.34	−0.13	−0.37	−0.62	−0.13
JD-Next MC only	65	0.51	0.39	192	0.65	0.35	−0.14	−0.41	−0.65	−0.16

Note. MC = multiple-choice.

intercepts showed a borderline significant difference in intercepts between the two groups, $p = .07$. Further investigation showed that this difference is based on a significant difference in means between the two groups in the outcome variable of law school GPA, $p < .01$, and not JD-Next exam, $p = .43$, which can be considered evidence of potential criterion bias, not test bias, and thus does not preclude using the examination (Meade & Fetzer, 2009).

In summary, the JD-Next overall score also showed consistently smaller score disparities for underrepresented students compared to the LSAT, as well as positive predictive validity for underrepresented test takers.

Score Disparities Based on Parental Education

Similar to analyses for the JD-Next MC score, we compared subgroup differences in the JD-Next overall scores to the LSAT scores (including GRE scores converted to LSAT scores) based on whether students were from households in which they were the first generation completing a bachelor's degree. Table A6 displays Cohen's d statistics for the JD-Next overall score, along with the JD-Next MC score, with sample sizes and 95% confidence intervals that account for unavoidable uncertainty in estimating effect sizes.

We also examined the predictive validities of both tests for first-generation test takers separately. Table A7 gives R^2 values for models including UGPA, LSAT (including converted GRE), and school median LSAT (including converted GRE).

The JD-Next exam has consistent incremental predictive validity across both groups. Also, a model adding a first-generation status indicator variable using the complete sample was fit to investigate potential differential prediction (i.e., predictive bias, differences in slopes and intercepts; Cleary, 1968). This model showed that the interaction term between the first-generation status indicator and JD-Next was not significant, $p = .52$, meaning that there was no significant

Table A7 Predictive Validity for First-Generation Students and Other Students

	First generation[a]			Other students[b]		
Variable(s) entered	R	R^2	ΔR^2	R	R^2	ΔR^2
Model 1						
UGPA and mLSAT	.32	.10		.35	.12	
JD-Next	.51	.26	.16**	.47	.22	.10**
Model 2						
UGPA and mLSAT	.32	.10		.35	.12	
LSAT	.60	.37	.26**	.51	.26	.13**
Model 3						
UGPA and mLSAT	.32	.10		.35	.12	
JD-Next	.51	.26	.16**	.47	.22	.10**
LSAT	.65	.42	.15**	.56	.31	.09**
Model 4						
UGPA and mLSAT	.32	.10		.35	.12	
LSAT	.60	.37	.26**	.51	.26	.13**
JD-Next	.65	.42	.05*	.56	.31	.05**

Note. LSAT = Law School Admission Test. mLSAT = median Law School Admission Test score for the participant's law school. UGPA = undergraduate grade point average. The LSAT variable includes converted GRE scores. [a] $n = 60$. [b] $n = 178$. *$p < .05$. **$p < .01$.

difference in slopes. Additionally, an exploration of the differences between intercepts showed no difference in intercepts between the two groups, $p = .47$. On the basis of this finding, no differential prediction was found.

In summary, the JD-Next overall score also showed consistently smaller score disparities for first-generation students compared to the LSAT, as well as positive predictive validity for first-generation students.

Suggested citation:

Holtzman, S., Steinberg, J., Weeks, J., Robertson, C., Findley, J., & Klieger, D. (2024). *Validity, reliability, and fairness evidence for the JD-Next exam* (Research Report No. RR-24-04). ETS. https://doi.org/10.1002/ets2.212378

Action Editor: Michael Walker

Reviewers: Harrison Kell and Margarita Olivera Aguilar