NextGen Research Brief: Field Test

October 11, 2024

www.ncbex.org

NCBE National Conference of Bar Examiners

Building a competent, ethical, and diverse legal profession.

National Conference of Bar Examiners

Building a competent, ethical, and diverse legal profession.

NextGen Research Brief: Field Test

Table of Contents

Introduction

In early 2021, following a three-year study focused on the current bar examination and the evolving landscapes of both the legal profession and legal education, the National Conference of Bar Examiners (NCBE) began to develop and implement the next generation of the bar exam (NextGen bar exam). This work includes multiple implementation research phases to test new question types and content and to establish administration and scoring processes prior to the NextGen bar exam's debut in July 2026. The purposes of each research phase are:

Implementation Research Phase 1: Pilot Testing

- Determine efficacy of new question types
- Determine impact of providing legal resources
- Determine time needed to answer new question types

Implementation Research Phase 2: Field Testing

- Gather data on additional question types
- Confirm timing estimates
- Compare grading experiences for the NextGen bar exam and the current exam
- Generate initial question and test performance data

Implementation Research Phase 3: Prototype Exam Administration

- Test jurisdiction administration of a full-length, nine-hour examination via the delivery platform that will be used for the live exam
- Test the written-response grading system that will be used for the live exam
- Generate performance data to set the new score scale, establish concordance between the current score scale and the new scale, and provide information for jurisdictions' determination of passing scores on the NextGen bar exam

Earlier this year, NCBE published a research brief describing findings from the first research phase, pilot testing. This field test research brief summarizes the second research phase, which focused on administering NextGen bar exam questions to a large sample of current law students and recent law graduates to determine the feasibility of the question formats and to conduct an initial evaluation of question-level performance. The brief begins with background information on the question types tested. It then provides a summary of the characteristics of the field test administrations, followed by discussion of the lessons learned and how they will inform prototype testing.

Questions that field testing sought to answer included:

1. Is it feasible to administer and score the proposed new question types?
2. How much time does it take participants to respond to the different question types?
3. How does the experience of grading the new exam compare to the experience of grading the current exam?
4. Do the questions reduce performance differences that are not related to examinee competency?

Field Test Question Types

Three types of questions were included on the field test:

Multiple-Choice Questions: Standalone multiple-choice questions with either four answer options and one correct answer or six answer options and two correct answers. Note that some multiple-choice questions also appeared in the integrated question sets and longer performance tasks described below. **Sample Multiple-Choice Questions**

Integrated Question Sets: Each set was based on a common fact scenario and could include some legal resources (e.g., excerpts of statutes or judicial opinions) and/or supplemental documents (e.g., a police report or excerpt from a deposition). Integrated question sets included a mixture of multiple-choice and short written-response questions. In addition to testing doctrinal law, some integrated question sets focused on drafting or editing a legal document; other sets focused on counseling and/or dispute resolution.
Sample Integrated Question Sets

Performance Tasks: These tasks required participants to demonstrate their ability to use fundamental lawyering skills in realistic situations, completing tasks that a beginning lawyer should be able to accomplish. These tasks could feature areas of doctrinal law, with accompanying legal resources, not included in the NextGen Foundational Concepts and Principles. One of the longer performance tasks included multiple-choice questions and short written-response questions focused on research skills, followed by a longer writing assignment. **Sample Performance Task**

Field Test Administration Characteristics

The field test was administered on January 26 and 27, 2024, to 4,016 final-year law students and recent law graduates at 88 volunteer law schools across the United States. Participants completed a collection of field test questions (a "form") that consisted of two hours of test content. By randomly assigning participants to different field test forms, we ensured that performance differences between forms could be interpreted as differences in form difficulty, not participant ability. Participants were compensated for their time.

Additional information regarding question types, participants, and forms is included below. The analyses conducted include question difficulty, response times, grader performance, and group performance by question type and skill category.

Field Test Questions and Forms

A total of five field test forms were administered. Across the five forms, there were 112 standalone single-selection multiple-choice questions. Additionally, four drafting sets assessing writing and editing skills not easily covered by the longer written portions of the exam were included. Five counseling sets assessing client-counseling, advising, negotiation, and dispute-resolution skills were included,[1] as were two performance tasks and one research performance task.[2] In total, 155 individual questions were administered.

All Foundational Concepts and Principles and Foundational Skills were tested across the 155 questions.[3]

How Did Field Test Questions Perform?

To examine question performance, the proportion of possible points earned across participants (p-value) was calculated. A higher p-value indicates that the question is easier (i.e., participants tended to earn more points), and a lower p-value indicates that the question is more difficult (i.e., participants tended to earn fewer points). Summary statistics of p-values for each question type are presented in Table 1.

---

[1] Each counseling set consisted of four short written-response questions and two single-selection and/or multiple-selection multiple-choice questions.

[2] This research performance task consisted of two multiple-selection multiple-choice questions, two single-selection multiple-choice questions, and four written-response questions.

[3] The Foundational Concepts and Principles and Foundational Skills can be seen at https://www.ncbex.org/exams/nextgen/content-scope.

Table 1. Summary Statistics of P-Values by Question Type

Question Type	Number of Questions	Mean P-value	SD	Minimum	Maximum
Written-response questions					
Drafting sets	4	0.63	0.12	0.48	0.76
Medium-length writing	1	0.56	—	0.56	0.56
Short answers	21	0.51	0.14	0.21	0.80
Performance tasks	3	0.60	0.06	0.53	0.64
Multiple-choice questions					
Multiple-select	7	0.65	0.14	0.47	0.91
Single-select	118	0.48	0.20	0.07	0.96

P-values can range between 0.0 and 1.0; in general, we desire questions with p-values in the middle of this range (close to 0.5) and want to avoid questions with p-values that are extremely low (less than 0.2) or extremely high (greater than 0.9).[4] Average field test p-values ranged between 0.48 and 0.65 across question types. Single-select multiple-choice questions showed the lowest average p-value (0.48), suggesting that, on average, these questions were more difficult than other question types. Also, though, compared to other question types, a wider range of p-values was observed for single-select multiple-choice questions (0.07 to 0.96), and more such questions were included on field test forms. P-values for written-response questions and multiple-select multiple-choice questions were also within reasonable ranges (0.21 to 0.80).

Some of the multiple-choice questions used on the field test also appeared on the February 2024 Multistate Bar Exam (MBE). Most of these questions had lower p-values (appeared more difficult) when administered during the field test than they did when they were used as part of the MBE. There could be several reasons for this difference: for example, field test participants might have been less motivated or less prepared than examinees taking the February bar exam. To determine whether field test participants may have been unmotivated, we analyzed field test questions of all types that had short response times and omitted responses. These analyses did not indicate concerning levels of lack of motivation from field test participants. We also compared the scores of law student participants with those of recent law graduates. The law students in the participant group had lower average scores than the recent graduates.

[4] American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (AERA, 2014), Standard 4.10.

These results provide a wealth of useful data consistent with preliminary findings from pilot testing that NextGen questions have an appropriate level of difficulty. At the same time, differences between field test performance and MBE performance are a reminder that the field test does not perfectly predict how questions will perform on a live exam. Additional testing and analysis during the prototype exam stage of implementation research, as well as statistical analysis of question performance on the live NextGen bar exam, will help ensure questions have the right level of difficulty.

How Much Time Does It Take Participants to Respond to the Different Question Types?

The exam delivery platform collected data on time spent responding to field test questions. On average, participants spent 1.3 minutes on each multiple-choice question, 17.6 minutes on each drafting set, 15.1 minutes on each counseling set, and 42.5 minutes on each performance task. To examine timing estimates, we also calculated the 90th percentile of time spent (the time within which 90% of the participants had completed their responses).[5] The 90th-percentile response time for multiple-choice questions was 2 minutes, which was slightly longer than the expected 1.8 minutes. The 90th-percentile response time for a drafting set was 25.9 minutes, roughly 2 minutes longer than the expected 24 minutes, and the 90th-percentile response time for a counseling set was 21.9 minutes, about 2 minutes less than the expected 24 minutes. Among the question sets, the longest average response time was 21 minutes; the 90th-percentile response time was 31.7 minutes. The 90th-percentile response time for a performance task or research performance task was 58.1 minutes, about 2 minutes less than the expected time of 60 minutes. A summary of this information is presented in Table 2.

Table 2. Summary of Average and 90th-Percentile Times by Question Type

Question Type	Average Time	90th-Percentile Time
Drafting sets	17.6 minutes	25.9 minutes
Counseling sets	15.1 minutes	21.9 minutes
Performance tasks	42.5 minutes	58.1 minutes
Multiple-choice questions	1.3 minutes	2 minutes

[5] AERA, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, Standard 3.1.

These results provide useful information that builds on the data gathered during the pilot testing phase. They will help us refine and finalize timing allotments for the live exam, ensuring that the exam will not be speeded for sufficiently prepared candidates.

How Does the Experience of Grading the New Exam Compare to the Experience of Grading the Current Exam?

The most significant change to grading from the current Uniform Bar Exam (UBE) to the NextGen bar exam may be the switch from today's relative grading model, which is applied within each jurisdiction independently (albeit with standardized grading materials), to an objective, absolute grading model applied across all jurisdictions. Both are fair, psychometrically valid approaches, but absolute grading rubrics are significantly more straightforward for graders. Absolute grading will benefit all jurisdictions, especially medium and large ones, which currently must maintain relative grading standards across a large volume of written responses. Evaluating absolute grading was a key component of the field test. The field test provided NCBE with the opportunity to examine how this change might affect graders.

Grading took place in February 2024. Sixty-one volunteers from 27 jurisdictions graded 37,000 written responses. Volunteers were graders for the February 2024 bar exam who jurisdiction administrators recommended for participation in field test grading. Grading assignments contained four counseling sets, two drafting sets, and one performance task. Graders were given the same number of questions per question type, but the questions themselves were different. Graders were asked to submit timing data as they were grading to give insight into how long it takes graders to review the scoring guides and grade an assigned question set. On average, graders spent 13 hours and 43 minutes grading their field test assignment. Two graders graded 11,127 (44%) of the 37,000 total responses (double grading), and 3,378 (30%) of those responses received additional review before a final grade was assigned (adjudication).

Graders participated in interviews after grading field test responses. Because they also graded the February bar exam for their jurisdictions, they were asked to provide feedback about the different approaches to grading (i.e., relative vs. absolute). Results suggest that the structured nature of the NextGen bar exam grading materials allowed graders to better align participant answers with the grading criteria, potentially leading to fairer and more accurate assessments. The more flexible approach in use for the current exam allows for grader discretion but may result in less consistency and clarity in evaluating examinee performance. Additionally, although the NextGen bar exam's detailed rubric and extensive supplemental materials enhanced consistency and clarity in grading, they initially presented challenges in terms of the time and effort required from graders.

For this reason, it is especially important that the prototype exam will provide jurisdiction graders with an opportunity to familiarize themselves with the new grading materials and processes before the first NextGen bar exam administration. Graders' timing data continues to support a key finding from the pilot test: once graders are acclimated to the NextGen question types and absolute grading methods, double-grading the NextGen bar exam is expected to take the same number of grader hours as single-grading the current UBE.[6]

Do the Questions Reduce Performance Differences That Are Not Related to Examinee Competency?

In the field test administrations, we were able to examine group performance by Foundational Skill. To do so, we obtained scores by Foundational Skill and calculated the percentage of total possible points received. This provided a way to make comparisons for different skill components. Descriptive statistics (mean and standard deviation) were generated, and mean comparisons were made based on demographic variables (law school year, first in family to earn a bachelor's degree, first in family to attend law school, whether participants identify as having a disability, whether English was their first language, race/ethnicity, and gender identity).

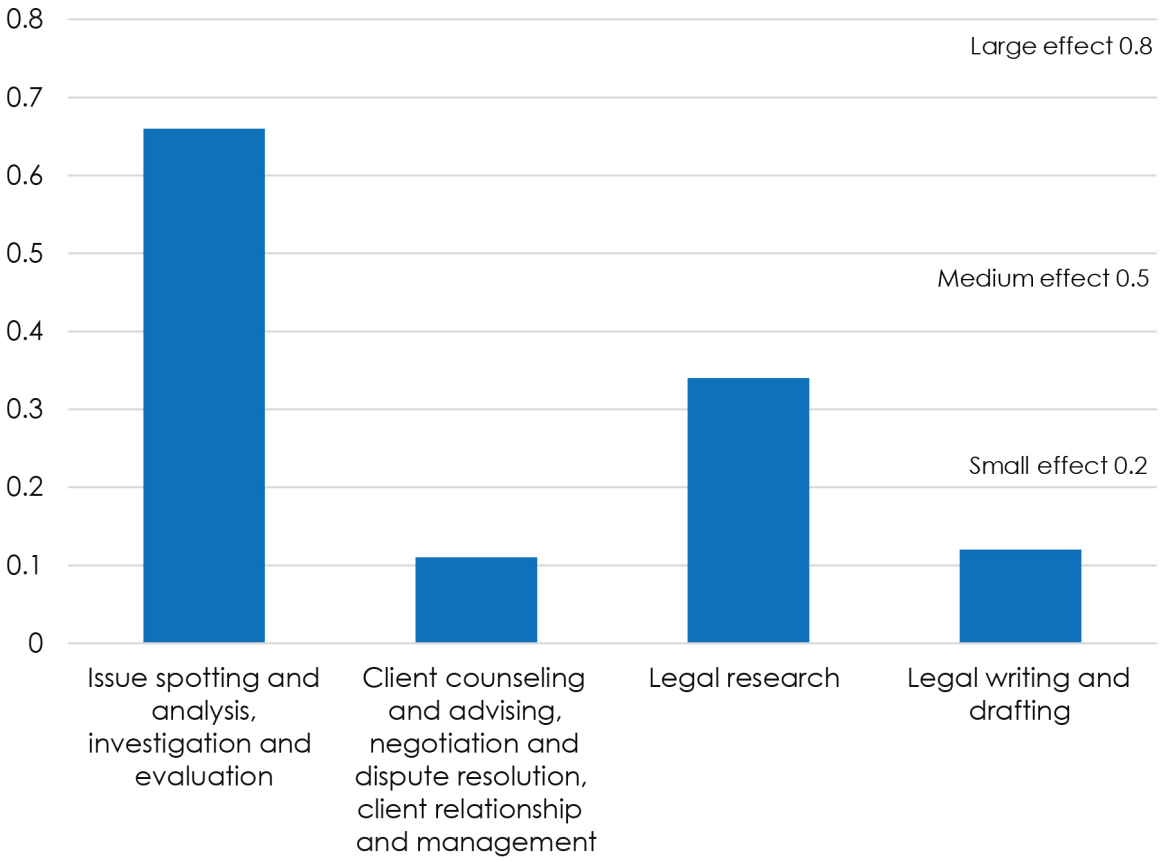
Standardized mean differences (d) were calculated to evaluate the magnitude of the effect (effect size d) of these demographic variables on performance. Effect size d is expressed in standard-deviation units to facilitate interpretation. To calculate effect size d, the mean score of the comparison group was subtracted from the mean score of the reference group and the difference was divided by the pooled standard deviation:

$$d = \frac{m_{reference} - m_{comparison}}{sd_{pooled}}$$

A positive value for effect size d indicates that the variable in question is associated with a higher proportion of possible scores of the reference group, on average, whereas a negative value indicates that the variable in question is associated with a higher proportion of possible scores of the comparison group, on average. We found that there were differences in performance based on year in law school (final year vs. recent graduates) by Foundational Skill; these differences are shown in Figure 1.

[6] Wendy Light; Rosemary Reshetar, EdD; and Erica Shoemaker, "The Testing Column: Grading the MEE, MPT, and the NextGen Bar Exam: Ensuring Fairness to Candidates," 93(1) *The Bar Examiner* 69–72 (Spring 2024), *available at* https://thebarexaminer.ncbex.org/article/spring-2024/the-testing-column-spring24/.

Figure 1. Standardized Mean Differences by Foundational Skill between Final-Year Law Students and Recent Graduates



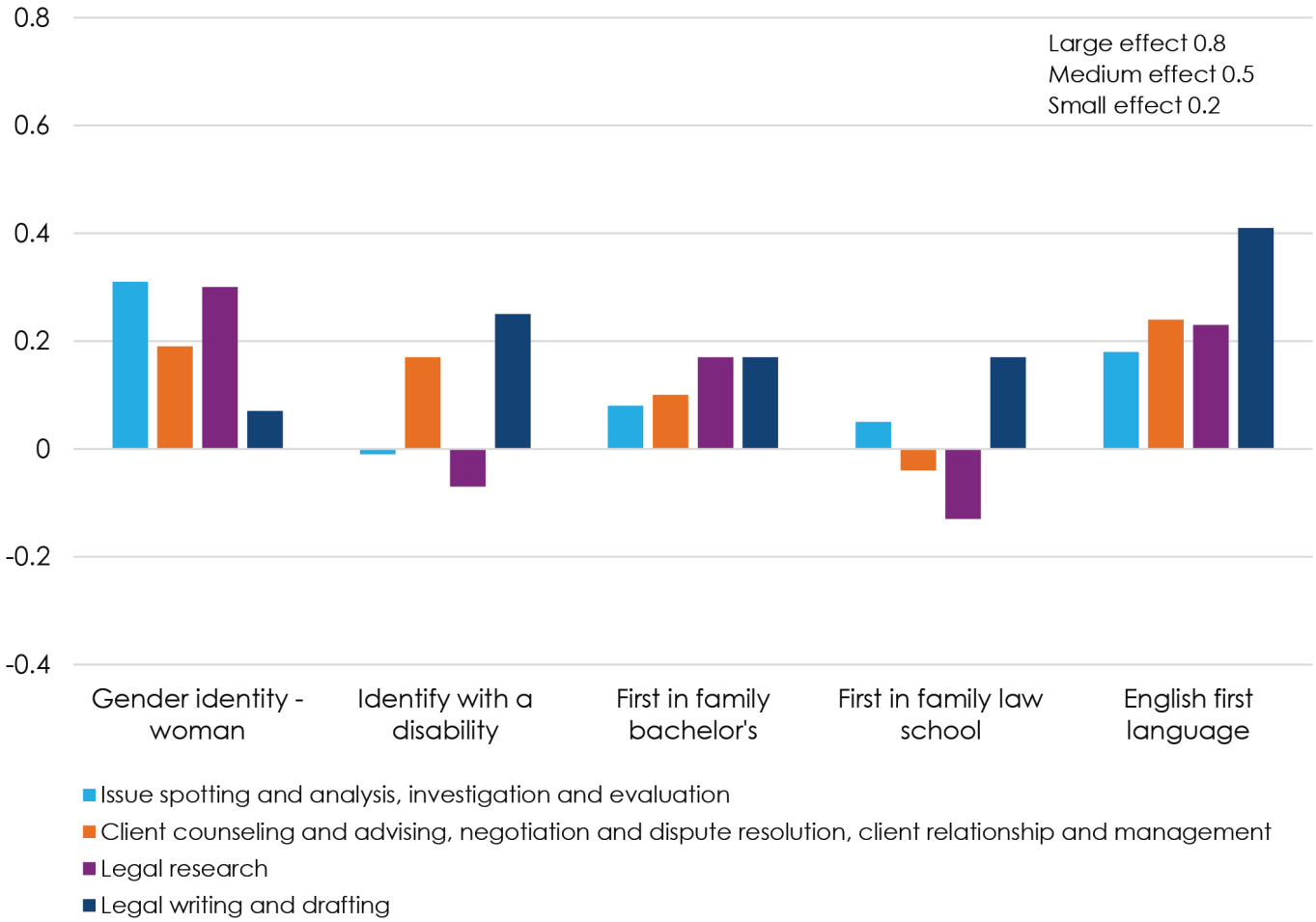
Recent graduates outperformed final-year law students on all skills. Using the standardized mean difference to illustrate the association between year in law school and field test performance, we found that performance differences were most pronounced for issue spotting and analysis, investigation, and evaluation (Group A; 53.9% for recent graduates vs. 44.5% for final year graduates) and legal research (Group C; 58.1% vs. 49.3%). The standardized mean difference for Group A skills was 0.67, a medium effect;[7] for Group C skills the standardized mean difference was 0.34, a small effect. Differences in scores for client counseling and advising, negotiation and dispute resolution, and client relationship and management (Group B; 54.2% vs. 51.3%), and for legal writing and drafting (Group D; 64.3% vs. 61.5%), were even smaller;[8] $d = 0.12$ for Group B and $d = 0.12$ for Group D. Because

[7] To interpret the magnitude of the effect sizes, we used commonly accepted categories: 0.01 to 0.49 indicated a small effect; 0.5 to 0.79 indicated a medium effect; .0.8 and higher indicated a large effect. Gail M. Sullivan and Richard Feinn, "Using Effect Size—or Why the P Value Is Not Enough," 4(3) *Journal of Graduate Medical Education* 279–282 (September 2012), *available at* https://doi.org/10.4300/jgme-d-12-00156.1.

[8] *Ibid.*

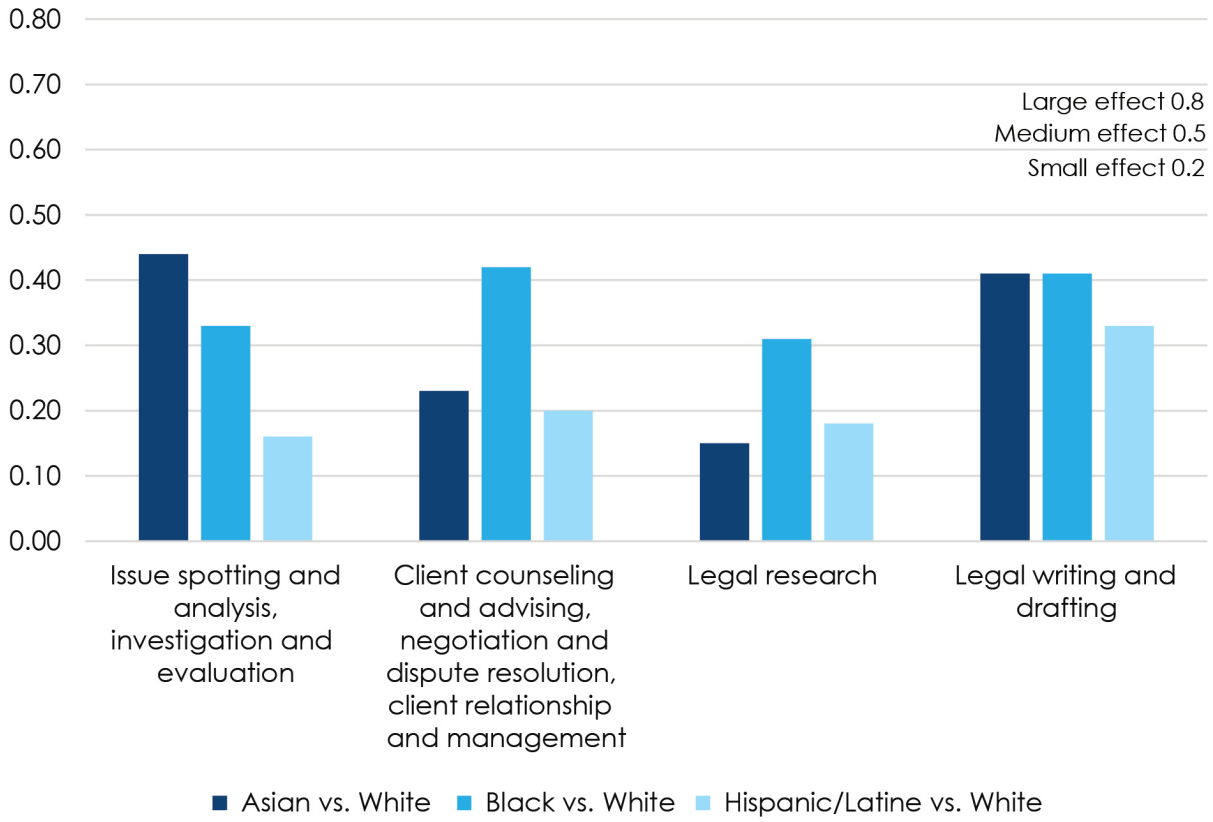
there were effects (though generally small) based on this variable, we focused the rest of our analysis on recent graduates, the group most similar to bar exam examinees. Those differences by Foundational Skills are discussed below.

Figure 2. Standardized Mean Differences for Select Demographic Variables



Standardized mean differences for demographic variables separate from race/ethnicity are provided in Figure 2. Analyses of recent graduates' performance by Foundational Skill to examine the effect of examinee characteristics showed that performance differences were small for most variables; standardized mean differences in scores ranged from −0.13, when comparing the performance on legal research of those who were first in family to attend law school to those who were not, to 0.41, when comparing the performance on legal writing and drafting of those who reported that English was their first language to non-native English speakers.

Figure 3. Standardized Mean Differences by Foundational Skill by Race/Ethnicity



The association of race/ethnicity with exam performance was considered as well. For this analysis, groups were categorized based on self-reported race/ethnicity. Groups with fewer than 10 members were not included. The proportion of points earned by Foundational Skill was compared for each group, using participants who reported they were White as the reference. The standardized mean differences ranged from 0.15 (Asian participants vs. White participants; legal research) to 0.44 (Asian vs. White participants; issue spotting and analysis). For this demographic variable, all effect sizes would be considered small.

Group performance differences by Foundational Skill were found in this analysis; however, all the differences are considered small effects. These differences should not be used to predict passing rates, as those rates also depend on the passing scores that jurisdictions will set. Analyses to assess whether new question types and test content result in a reduction in the magnitude of the difference in scores will be completed after the prototype research phase, which will also include generating recommendations for passing scores.

Conclusion

This report provides an overview of the field test phase that was completed as part of NextGen bar exam research and development work. Key findings from this research phase can be summarized as follows:

1. **How did field test questions perform?** Analyses of questions answered in the field test suggest that they had average difficulty and showed insignificant differential performance by subgroup.

2. **How much time does it take participants to respond to the different question types?** On average, participants spent 1.3 minutes on multiple-choice questions, 17.6 minutes on drafting question sets, 15.1 minutes on counseling sets, and 42.5 minutes on performance tasks.

3. **How does the experience of grading the new exam compare to the experience of grading the current exam?** The extensive grading materials provided to graders initially required more time and effort compared to current grading practices. However, the structured nature of the materials allowed graders to align participant responses with grading criteria, potentially leading to fairer and more accurate assessments. Once graders are oriented to the new exam content and grading rubrics, double-grading the NextGen bar exam is expected to take roughly the same amount of time as single-grading the UBE.

4. **Do the questions reduce performance differences that do not reflect examinee competency?** Group differences based on participant demographics and Foundational Skills were found among recent graduates; however, all the differences are considered small effects. Analyses to assess whether new question types and test content result in a reduction in the magnitude of the difference in scores will be completed after the prototype research phase.

The prototype exam phase of implementation research will build on these findings to confirm performance and timing data in support of test design, including the generation of a new score scale for score reporting and passing score decisions. Additionally, the prototype exam will provide experience with NextGen systems and processes for jurisdiction administrators and graders.