NextGen Research Brief: Pilot Testing

May 28, 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: Pilot Testing

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. That work, which will culminate in the first administration of the NextGen bar exam in July 2026, includes multiple stages of implementation research to test new question types and content and to establish administration and scoring processes.[1] These stages are:

Pilot Testing: Small-scale administration of drafts of new question formats under semirealistic conditions to groups of law students and recently licensed practicing lawyers.

Field Testing: Large-scale administration of finalized new questions under realistic conditions to a large representative sample of students and recently licensed lawyers for the purpose of determining the operational feasibility of the questions and format and obtaining estimates of question performance statistics.

The results described in this research brief involve several different question types:

Standalone Question: This question type is a selected-response (aka multiple-choice) question. Examinees must pick one correct answer from among four options or two correct answers from among six options.

Question Set: Factual and legal information pertaining to a common scenario is presented to the examinee, who must answer multiple questions about the scenario. Sometimes, additional materials are introduced as the examinee progresses through the question set. The questions are either selected-response or short constructed-response questions, where the examinee must give an answer ranging from a sentence to a short paragraph in length. This is also described in NextGen materials as an Integrated Question Set.

Longer-Form Writing: The examinee is presented with factual and legal materials and must write a long answer in the form of a memorandum, legal brief, letter, or similar document. Occasionally, longer-form questions are preceded by a few selected-response questions designed to assess skills specific to legal research. This question type was not part of the pilot test but is included in the subsequent field test and prototype exam. This is also described in NextGen materials as a Performance Task.

For more information about these question types and to see sample questions, visit https://www.ncbex.org/exams/nextgen/sample-questions.

1 For a timeline of past and upcoming research phases, visit https://www.ncbex.org/exams/nextgen-july-2026/about/implementation-timeline.

Prototype Exam: A large-scale administration of a full-length exam to a representative sample of recent examinees for the purpose of generating performance data to support the standard-setting process for jurisdictions.

This brief summarizes the first research phase, which focused on pilot testing new question types, including those measuring a broader set of lawyering skills. The brief begins by providing background information on recommendations made by NCBE's Testing Task Force, which has guided development of the initial NextGen bar exam content and design. It then provides a summary of the characteristics of the pilot test administrations, followed by discussion of lessons learned from those administrations.

Subsequent research phases have already begun: the first field test, which focused on further development of exam questions and on exam delivery, was administered in January 2024. A full-length prototype test, which will mimic the real NextGen exam, will be administered in October 2024. NCBE will issue research briefs summarizing these phases after they are complete.

Background

In 2018, NCBE appointed a Testing Task Force charged with undertaking a three-year study to ensure that the bar examination continues to test the knowledge, skills, and abilities required for competent entry-level legal practice in a changing profession. The study's primary goal was to identify the foundational knowledge and skills that should be included on the NextGen bar exam and to determine how they should be assessed. The final recommendations of the Testing Task Force included using an integrated exam structure with a combination of selected- and constructed-response question formats covering essential knowledge and skills areas (referred to as Foundational Concepts and Principles and Foundational Skills). More detailed information about the Testing Task Force's final recommendations and the rationale for its decisions is available in the report found here: https://nextgenbarexam.ncbex.org/overview-of-recommendations/.

To begin development of the new exam, the Testing Task Force recommended that NCBE conduct research on the feasibility of developing new question types, on the use of a computer-based examination format, and on the determination of a reasonable time required to answer the various question types. Additional validity evidence was to be collected on using legal resources provided during the NextGen bar exam, understanding cognitive processes of pilot test participants, and comprehending differences in performance based on various demographic and other contextual characteristics. To gather this data, NCBE conducted pilot testing in 2022 and 2023.

Characteristics of Pilot Test Administrations

From August 2022 to April 2023, NCBE conducted four administrations of pilot testing to evaluate new questions and question sets (groups of questions associated with an overarching scenario). These questions and question sets varied in several ways, including:

- format (e.g., single-selection or multiple-selection selected-response questions, short-response questions, extended-length constructed-response questions);
- the number of total options and keys (correct answers) in selected-response questions;
- the composition of question sets (e.g., multiple selected-response and short-response questions that pertain to a common stimulus, such as a fact pattern); and
- grading scales for constructed-response questions.

Multiple pilot test forms consisting of groups of questions and question sets were administered to over 2,500 participants (primarily final-year law students and recent law school graduates) using their own devices. The sample size in these administrations varied, with as many as 711 participants per pilot test form.[2] NCBE made efforts to recruit participants from law schools in different regions and with different student-body profiles to obtain diverse participant groups for the pilot testing. There were, however, some imbalances in subgroup representations. For instance, the number of female participants was approximately twice that of male participants. This contrasts with a typical bar exam population, which usually has similar numbers of male and female examinees.

Data including participants' responses to the questions and question sets and their response times were collected and analyzed. Participants' demographic information and their feedback on each collection of questions or question sets, the amount of time needed to complete the questions and question sets, and their overall pilot testing experience was also gathered using NBCE-developed survey questions.

Pilot testing sought to answer the following questions:

1. Is it feasible to administer and score the proposed new question types?
2. What is the effect of providing legal resources for the new question types?
3. How much time does it take participants to respond to the different question types?

2 Note: More than one pilot test form might be administered in a single event/administration.

4. What cognitive processes do study participants use in responding to the different question types?
5. Do the new question types reduce performance differences that do not reflect examinee competency?

Is It Feasible to Administer and Score the Proposed New Question Types?

Pilot test participants generally found the NextGen bar exam question formats to be practical, engaging, and more skills-oriented than those used on the current bar exam. Participants did raise concerns about ambiguity in some questions and instructions, leading the test development team to improve question quality via evaluation and revision.

To ensure that the NextGen bar exam enables jurisdictions to make reliable decisions about examinees' readiness to practice law, questions on the exam must be neither too easy nor too difficult; both very easy and very difficult questions do not contribute to those decisions because even the most seasoned lawyers can get very hard questions wrong while even underprepared candidates can get very easy questions right. Based on feedback on survey questions, most pilot test participants (between 92% and 100%) indicated that they found the question sets and standalone questions to be of reasonable difficulty (i.e., neither extremely easy nor extremely difficult). For most questions, participants' perceptions of difficulty were consistent with the proportion of participants who answered the questions correctly (i.e., classical item difficulty). For a few selected-response questions, the proportion of participants answering the questions correctly was very high or very low, suggesting that those questions might be overly easy or difficult and might require further refinement to ensure they meet the expected difficulty levels.

The process of developing and revising grading rubrics for constructed-response questions, as well as the process for grader training, monitoring, and grading, were refined and standardized through the pilot test administrations. Refinement and standardization of the rubrics and grading process improved reliability in grading, which was reflected in better usage of score points along the grading scales and increased grader agreements and Cohen's kappa coefficients (a statistical measure of agreement across graders).

What Is the Effect of Providing Legal Resources for the New Question Types?

Early in the test development process for the NextGen bar exam, a Test Design Committee convened to address specific design recommendations from the Testing Task Force. One question the committee raised was whether access to the complete or abridged Federal Rules of Evidence (FRE) or other standard legal resources should be provided to examinees during the NextGen bar exam. To address this question, a pilot test form was administered to two randomly assigned participant groups. One group received instructions and a link to access and use the FRE during the pilot test; the other group was not given such access. The demographic distributions for the two groups (FRE and non-FRE) were very similar, enabling a reasonable comparison of performance and response time between the two groups.

Of the participants with access to the FRE (the FRE group), 72% reported using the FRE during the exam (the FRE user subgroup), and the remaining 28% reported not using them. Table 1 below shows summary statistics regarding the proportion of points earned by the FRE group, the FRE user subgroup, and the non-FRE group. The results show minimal difference in performance at the pilot test form level across the three different groups of participants. In addition, FRE participants generally spent more time than non-FRE participants answering questions requiring FRE knowledge, suggesting that access to such legal resources slows response time without increasing examinee scores. Participants also indicated that they used the FRE materials mostly for checking, rather than composing, their answers. These findings supported the decision to exclude access to additional legal resources from the NextGen bar exam.

Table 1. Summary Statistics of Proportion of Points Earned by Each Participant Group/Subgroup

Group/Subgroup	Mean Proportion of Points Earned	Standard Deviation	Minimum	Maximum
Non-FRE (N = 698)	0.62	0.17	0.20	0.85
FRE (N = 711)	0.61	0.16	0.22	0.88
FRE User (N = 514)	0.62	0.16	0.22	0.88

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

The observed response time to pilot test questions varied across different question types, as expected. On average, participants spent around 1.1 minutes on a selected-response question, approximately 4 minutes on a short-answer question, and about 20 minutes on a question set. In addition, a supermajority of participants completed the questions just within the time we estimated or faster, ensuring that the exam will not be speeded for sufficiently prepared candidates. This timing pattern closely aligned with participants' perception of the amount of time they needed to answer the questions. It should be noted that in the pilot test administrations, participants effectively had as much time as they wanted to respond to questions and were allowed to work at their preferred pace. This differs from actual exam circumstances. Moreover, participants' motivation may have been lower than if they were taking a full, operational bar exam. However, the timing results provided valuable insights for determining the appropriate time allotment for questions and question sets, as well as for optimizing the composition of question types within the NextGen bar exam form. The field test phase will support further analysis of timing for all question types.

What Cognitive Processes Do Study Participants Use in Responding to Different Question Types?

NCBE conducted a think-aloud study in which 25 participants were asked to verbalize their thought processes as they completed a group of NextGen questions and question sets. This study sought to investigate the cognitive processes different examinees use and relate that to exam performance.

As noted above, the study found that participants had positive responses to the new NextGen question types and test format. Participants especially noted that the questions had more relevance to their professional lives, both as early-career attorneys and interns, and better tested skills that practicing attorneys use.

The study also found a differentiation between cognitive processes used for different question types. Participants relied on legal reasoning and knowledge, particularly knowledge acquired in clinics, internships, and classrooms, while completing the question sets. Participants also relied on legal knowledge while completing standalone selected-response questions; however, they also relied on test-taking strategies to arrive at final answers (e.g., using the process of elimination, reading the question stem, etc.). The reliance on test-taking strategies may be due to the format of selected-response questions. The

analysis also identified interpretive issues in the questions, instructions, and fact-pattern language. Finally, participants noted practical problems with the test delivery platform and the use of FRE resources, including difficulty navigating between the FRE and exam pages. This finding solidified the decision to not include FRE on the final iteration of the exam.

Do the New Question Types Reduce Performance Differences that Do Not Reflect Examinee Competency?

Group differences in test scores that do not reflect differences in competency raise concerns about exam fairness. A group analysis study was conducted using data collected from pilot testing to compare participant performance across various subgroups. This helped determine whether the new question types affect previously observed group differences. The study supplements the analyses done to address the other research questions summarized above and provides recommendations for next steps in accumulating more validity evidence.

The results from the group analysis suggest that the new question types yield similar results across different groups of participants, reducing performance gaps between them. Although we found patterns of group differences in scores for factors such as disability status, exam accommodations status, non-native English-speaking status, race/ethnicity, and gender, the magnitude of these differences varied by group and question type. For example, differences between men and women were almost nonexistent for short-response questions. Differences based on being a non-native English speaker were considered small[3] and ranged from two-tenths of a standard deviation for single-selection selected-response questions to one-third of a standard deviation for multiple-selection selected-response questions and short-response questions. Performance differences were small for other variables as well; being first in one's family to receive a bachelor's degree or first in one's family to attend law school had no effect on performance for these question types. Though the pilot test data is encouraging, NCBE will continue to assess the interplay of question types and performance factors in subsequent research.

3 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.

Conclusion

This report provides an overview of the pilot testing study that was completed as part of NextGen bar exam research and development work. Key findings from this first empirical study can be summarized as follows:

1. **Is it feasible to administer and score the proposed new question types?** Yes, it is feasible to administer and score the new question formats, including question sets and multiple-select selected-response questions. For a more in-depth discussion, *see* https://thebarexaminer.ncbex.org/article/fall-2023/the-testing-column-fall23/.
2. **What is the effect of providing legal resources for these new question types?** Results supported the exclusion of external resources (e.g., the Federal Rules of Evidence).
3. **How much time does it take participants to respond to the different question types?** On average, participants spent around 1.1 minutes on a selected-response question, approximately 4 minutes on a short-answer question, and about 20 minutes on an entire question set.
4. **What cognitive processes are used by study participants in responding to the different question types?** Cognitive processes used to respond to new question formats are aligned with the skills identified in the practice analysis.
5. **Do the new question types reduce performance differences that do not reflect examinee competency?** Preliminary results indicate that differences in average scores across groups are small.