The NextGen Uniform Bar Examination

A Comprehensive Report

on Design, Development, and Delivery

May 2026

NCBE National Conference of Bar Examiners

Building a competent, ethical, and diverse legal profession.

NextGen UBE Uniform Bar Examination

The National Conference of Bar Examiners, founded in 1931, is a not-for-profit corporation that develops licensing tests for bar admission and provides character and fitness investigation services. NCBE also provides testing, research, and educational services to jurisdictions; provides services to bar applicants on behalf of jurisdictions; and acts as a national clearinghouse for information about the bar examination and bar admissions.

Our Mission

NCBE promotes fairness, integrity, and best practices in admission to the legal profession for the benefit and protection of the public. We serve admission authorities, courts, the legal education community, and candidates by providing high-quality

- assessment products, services, and research;
- character investigations; and
- informational and educational resources and programs.

Our Vision

A competent, ethical, and diverse legal profession.

NCBE
National Conference of Bar Examiners

Building a competent, ethical, and diverse legal profession.

NextGen UBE Uniform Bar Examination

302 South Bedford Street
Madison, WI 53703

ncbex.org
thebarexaminer.org
ncbex.org/exams/nextgen
nextgenbarexam.ncbex.org

/ncbexaminers
company/ncbex
/ncbexaminers

Acknowledgment of External Experts and Partners

NCBE extends its deepest gratitude to the more than 10,000 participants in NextGen UBE research studies, testing, surveys, and focus groups, as well as to the many jurisdiction representatives who provided essential insights and engagement. Their contributions provided the data, feedback, and perspectives that made the development of the NextGen UBE possible.

We also gratefully acknowledge the many experts, partners, and organizations whose collaboration supported this work, particularly:

- **The Center for Advanced Studies in Measurement and Assessment (CASMA)** – for conducting the concordance studies and providing independent validation of results.
- **The Human Resources Research Organization (HumRRO)** – for leading the May 2025 standard-setting workshop and ensuring best practices throughout the process.
- **Internet Testing Systems (ITS)** – for their partnership in developing and delivering the secure, reliable platforms used for exam delivery and constructed-response scoring.
- **Level Access** – for conducting a comprehensive third-party accessibility audit across the exam delivery system and all associated platforms to ensure full access for all examinees.
- **The NCBE Board of Trustees** (Hon. Scott Bales, Hon. Arlene Coleman Romeo, Timothy Davis, Judith Gundersen, John McAlary, Hon. Solomon Oliver, Jr., Hon. Shellie Park-Hoapili, Lisa Perlen, Augustin Rivera, Jr., Hon. Mary Russell, Darin Scheer, Anthony Simon, Hon. Ann Scott Timmer) – for their leadership and commitment to fairness and quality in licensure testing.
- **NCBE's Technical Advisory Panel** (Gregory J. Cizek, Deborah Harris, Carol Morrison, Mark Raymond) – for their independent review and expert guidance.
- **NCBE's Passing Score Advisory Panel** (Hon. Cynthia Martin, Augustin Rivera, Jr., Darin Scheer, Hon. Phyllis Thompson, Timothy Wong) – for their policy insight and jurisdictional perspectives.

Finally, NCBE extends special recognition to its staff, whose expertise and tireless work and communications made the NextGen UBE possible.

Contents

Contents (continued)

Executive Summary

Key Takeaways

- The NextGen Uniform Bar Examination is ready for operational launch.
- Psychometric performance is stable, reliable, and replicable across administrations.
- The recommended passing score range is supported by multiple lines of evidence.
- The six newly developed digital platforms and operational model function reliably under real-world conditions.
- The exam strengthens measurement of applied lawyering skills without reducing rigor.
- The system as a whole—content, platform, and scoring—works as intended.

The NextGen Uniform Bar Examination (NextGen UBE) is ready for operational launch.

This conclusion is based on a multi-year program of design, testing, and validation that evaluated not only exam content but the full assessment system under operational conditions.

At its core, the NextGen UBE is designed to measure what matters for modern legal practice: not only what examinees know, but what they can do with that knowledge.

The NextGen UBE is a deliberate, evidence-driven evolution of licensure—one designed to strengthen the connection between what is tested and what competent lawyering requires, while preserving the standards that protect the public.

Why Change Was Necessary

The structure of the bar exam has remained largely stable even as the practice of law has evolved. Entry-level lawyers today work in environments that are more integrated, more technology-enabled, and more dependent on applied reasoning than in prior decades.

Extensive engagement with stakeholders including courts, regulators, legal educators, and practitioners pointed to a consistent conclusion: doctrinal rigor must be preserved, but the exam must better assess how knowledge is used in practice.

The NextGen UBE was designed to meet that need. The objective was not to make the exam easier, but to make it more representative of the competencies required for safe and effective entry-level practice.

Three principles guided this work:

Relevance
Measure knowledge and skills demonstrably connected to entry-level practice.

Validity
Support score interpretations with empirical evidence.

Fairness
Minimize construct-irrelevant variance so performance reflects competence.

A Deliberate, Evidence-Driven Development Process

Because making changes to the bar exam introduces both psychometric and operational complexity, NCBE adopted a longitudinal, evidence-driven development approach. The exam was evaluated through a staged testing arc:

Pilot Testing (2022–2023)
Evaluation of early item families and scoring models

Field Test (January 2024)
Large-scale administration to assess item performance, timing, and grading feasibility

Prototype Exam (October 2024)
First full-form administration to establish scale, reliability, and standard setting

Beta Administration (January 2026)
Fully integrated administration under operational conditions

Each phase served a distinct purpose. Together, they produced a cumulative and converging body of evidence supporting both the design of the exam and its readiness for operational use.

Psychometric Evidence: Stability, Reliability, and Validity

Across administrations, the NextGen UBE has demonstrated stable and consistent psychometric performance.

- **Item Difficulty:** Distributions remain appropriately targeted and replicate across administrations.
- **Discrimination:** Items effectively differentiate among examinees, with no loss of precision following blueprint finalization.
- **Reliability:** Reliability estimates meet expectations for high-stakes licensure testing, supporting consistent decision-making at passing thresholds.
- **Dimensionality:** Analyses support interpretation of the exam as measuring a unified construct of minimum competence.
- **Fairness:** Differential item functioning analyses show no evidence of systemic bias across examined subgroups.

Most importantly, the January 2026 beta administration replicates the findings of the October 2024 prototype exam under fully operational conditions. This confirms that the measurement system performs as intended not only in controlled settings, but in live administration.

Validation of the Passing Score Range

The recommended passing score range was established during the prototype phase through multiple independent lines of evidence:

- standard-setting studies using expert judgment
- concordance analyses with the legacy UBE
- outcome-based validation across jurisdictions

The beta administration provides confirmatory evidence supporting this recommendation. Score distributions, reliability estimates, and standard error of measurement remained stable under operational conditions. No evidence emerged of systematic shifts in difficulty, discrimination, or subgroup performance that would call into question the reporting scale or passing standard.

Taken together, the evidence supports a clear conclusion: the recommended passing score range is valid, stable, and appropriate for operational use.

Operational Readiness: System Performance at Scale

The NextGen UBE is not only a new exam—it is a new digitally integrated assessment system. The beta administration evaluated that system end to end.

Delivery Platform

- Stable performance under live conditions
- Successful failover during connectivity interruptions with no loss of response data
- Continuous response capture supporting data integrity

Administrative Systems

- Real-time monitoring of candidate status across locations
- Structured incident classification and escalation
- Complete reconciliation of attendance and response records

Scoring and Grading

- Large-scale implementation of independent double grading
- Controlled reconciliation workflows
- Full auditability of scoring decisions

Beta completion rates exceeded 99% → with no evidence of system-related failures affecting examinee performance.

These results demonstrate that the operational model supports secure, consistent, and scalable administration.

An Integrated System

A defining feature of the NextGen UBE is the integration of its components into a single ecosystem. Candidate readiness, jurisdiction oversight, delivery, monitoring, and scoring operate through coordinated systems with clearly defined roles and controlled data flows.

This architecture enables

- separation of delivery, monitoring, and scoring functions;
- real-time administrative visibility without content exposure; and
- secure, auditable management of data across the exam lifecycle.

The result is not a digital version of a paper exam, but a structured system designed for modern licensure.

Conclusion: Evidence and Readiness

The evidence across design, testing, and operations is consistent.

The NextGen UBE demonstrates

- stable and replicable psychometric performance;
- valid and defensible passing-score recommendations;
- reliable and scalable scoring processes;
- secure and effective digital delivery; and
- operational readiness under real-world conditions.

With this evidence in place, the NextGen UBE is ready to support the jurisdictions and candidates NCBE serves.

Background and Introduction

The legal profession is changing.
NCBE has evolved licensure to change with it.

The NextGen UBE is designed to measure what matters for modern practice: not only what examinees know, but what they can do with that knowledge under realistic conditions.

Doctrinal knowledge remains essential. But entry-level lawyers are also expected to analyze complex problems, apply law in context, communicate clearly, and operate in digital, client-centered environments. The NextGen UBE assesses these capabilities together, producing a more complete and defensible measure of minimum competence.

This is not a simple modernization. Integrating skills with doctrine, delivering an assessment digitally, and scaling structured grading introduce real psychometric and operational complexity. Each design choice carries implications for validity, reliability, and fairness—and each has been tested through a deliberate, multi-phase process.

That process has been longitudinal and evidence-driven. Through pilot testing, field tests, a full prototype exam, and a large-scale beta administration, NCBE has evaluated both individual components and the performance of the full system under operational conditions.

Our guiding principle has remained constant: every change must preserve or strengthen the validity, reliability, and fairness that jurisdictions expect. This report documents how that standard has been met.

Part I. Origins and Foundations

Licensure Needs

The NextGen UBE was developed in response to evolving demands in legal practice and the enduring responsibility of licensure to protect the public. A licensure examination must measure minimum competence — the threshold knowledge and skills necessary for safe and effective entry-level practice.

Over time, courts, regulators, practitioners, legal educators, and psychometric experts recognized that the structure of the bar examination warranted modernization. Legal practice has become increasingly integrated, technology-mediated, and client-centered. Entry-level lawyers are expected to analyze problems holistically, apply doctrine in context, interpret foundational documents, and exercise professional judgment under conditions that more closely resemble practice.

The NextGen UBE was designed to align licensure assessment with contemporary entry-level practice while preserving its regulatory purpose. The objective was certainly not to lower standards, but to refine measurement — ensuring that the constructs assessed reflect competencies required for responsible lawyering.

Three licensure principles guided the redesign:

1. Relevance

The examination must measure knowledge and skills demonstrably connected to entry-level legal practice.

2. Validity

Score inferences must be supported by empirical evidence across content, structure, scoring, and outcomes.

3. Fairness

The assessment must minimize construct-irrelevant variance so that performance reflects competence rather than testing artifacts.

The development of the NextGen UBE represents a measured evolution of the general licensure exam, grounded in empirical study and regulatory purpose.

Testing Task Force and Stakeholder Engagement

To ensure that redesign decisions were evidence-based, transparent, and defensible, NCBE established a Testing Task Force to oversee a multi-phase program of inquiry. The Task Force employed a staged methodology—broad stakeholder listening, a national practice analysis, and targeted expert committees—so that content and structural decisions would rest on converging qualitative and quantitative evidence rather than assumption or preference.

PHASE 1:
Stakeholder Listening

The Task Force conducted extensive listening sessions with bar admission officials, legal educators, practitioners, judges, and other stakeholders. These sessions surfaced shared priorities, practical constraints, and areas of concern.

Recurring themes included strong support for increasing measurement of applied lawyering skills; concern that certain multiple-choice approaches may assess a level of knowledge beyond entry-level competence; and consistent emphasis on portability, fairness, and public protection. These perspectives informed both the scope and the direction of subsequent empirical inquiry.

PHASE 2:
National Practice Analysis

NCBE conducted a large-scale practice analysis survey to document the frequency and criticality of tasks, knowledge areas, and skills performed by newly licensed lawyers. The study produced a prioritized taxonomy of tasks and skill domains that can reasonably be expected of entry-level practitioners.

This analysis provided the empirical foundation for determining what a general licensure examination should measure. Content and skill domains included in the NextGen UBE Blueprint are directly traceable to this body of evidence.

PHASE 3:
Expert Panels

Focused expert committees translated stakeholder insights and practice-analysis findings into concrete recommendations regarding blueprint priorities, item family specifications, and scoring models. These panels addressed core design tradeoffs—such as the balance between doctrinal breadth and applied task performance—and proposed an integrated exam structure that preserves the testing of foundational legal knowledge while substantially increasing measurement of applied lawyering competencies.

Synthesis and Implications

The combined effect of stakeholder engagement, empirical practice analysis, and expert judgment established a documented chain of evidence linking licensure purpose to blueprint structure and item-design rules.

That chain of evidence is central to the exam's defensibility, jurisdictional adoption, and public legitimacy. The Testing Task Force's work provides both the mandate and the implementation framework for the NextGen UBE—defining not only what should be measured, but how those measurements must be constructed, scored, and validated.

A detailed report on the Testing Task Force's work is available at nextgenbarexam.ncbex.org/reports/final-report-of-the-ttf/.

Constructs, Blueprint, and Question Types

Construct Definition

The NextGen UBE is designed to support a validity argument regarding minimum competence for newly licensed lawyers. A newly licensed lawyer is defined as an individual within the first three years of practice. The target construct (i.e., what the exam is intended to measure) is the integrated ability to apply foundational legal principles through core lawyering skills under standardized, time-bound conditions.

The construct comprises two interdependent domains:

- **Foundational Concepts and Principles,** consisting of eight doctrinal areas: business associations and relationships, civil procedure, constitutional law, contracts, criminal law and constitutional protections of accused persons, evidence, real property, and torts
- **Foundational Lawyering Skills,** including issue spotting and analysis, investigation and evaluation, client counseling and advising, negotiation and dispute resolution, client relationship and management, legal research, and legal writing and drafting

These domains were derived from a national practice analysis documenting the frequency and criticality of tasks performed by entry-level lawyers. The examination does not treat knowledge and skill as independent constructs. Rather, it measures the coordinated application of doctrine within structured lawyering tasks.

Knowledge is assessed both independently and in context. Skills are assessed through both constrained, analytically scored tasks and extended written performance. The construct definition therefore establishes the first link in the evidentiary chain:

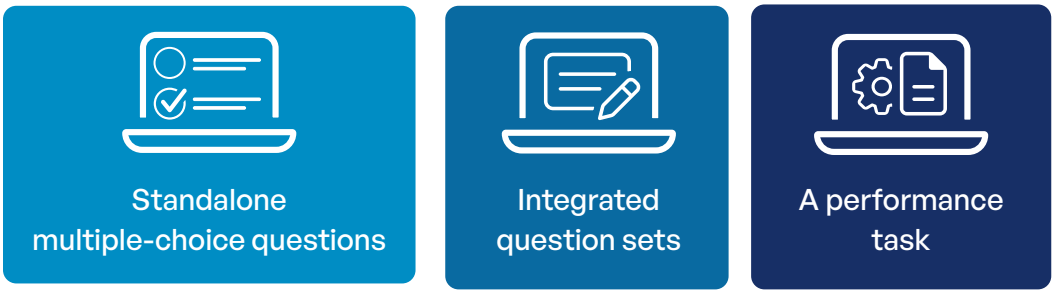
licensure purpose → domain specification → observable performance.

Full details about the NextGen UBE Content Scope are available at ncbex.org/exams/nextgen/content-scope.

Blueprint Architecture

The NextGen UBE Blueprint operationalizes the construct into enforceable test specifications. The examination consists of three three-hour sections administered over one and a half days.

Each section contains:



Across the full examination, score weighting is fixed at:

These proportions reflect deliberate design decisions concerning breadth and depth of measurement. Standalone multiple-choice questions permit broad sampling across doctrinal domains and provide psychometric stability for equating. Integrated question sets assess structured application of doctrine within contextualized client scenarios. Performance tasks require sustained integration of research, analysis, and written communication.

The Content Scope defines the boundaries of each doctrinal area and specifies how Foundational Skills are embedded within task structures. The starred and unstarred topic classification regulates whether legal resources are provided or whether examinees must rely on recalled knowledge. This distinction governs cognitive demand and preserves construct clarity.

Together, the Blueprint and Content Scope operate as design constraints. They

- regulate domain coverage;
- prevent construct drift across administrations;
- support form comparability; and
- establish structural preconditions for equating and score interpretation.

Item Family Specifications

Item families translate Blueprint specifications into replicable design templates. Each family defines

- targeted skill domain(s);
- permissible doctrinal scope;
- cognitive demand;
- response format;
- scoring architecture; and
- resource conditions.

This family-based architecture standardizes measurement while permitting content variation. It ensures that different forms represent the same underlying construct.

Standalone multiple-choice families are designed to sample doctrinal application efficiently. Two response formats are used: single select and multi select, with partial credit available in the latter. Items are independent and time-calibrated, supporting stable scaling and broad coverage.

Integrated question set families require examinees to complete structured lawyering tasks within a shared fact pattern. Drafting sets emphasize written analytical production. Counseling sets combine short-answer and multiple-choice components. These sets are analytically scored and constructed to elicit observable behaviors aligned to specified Foundational Skills.

Performance task families require sustained integration of knowledge and skills under realistic file-and-library conditions. Legal resources are embedded. Analytic rubrics allocate points across discrete performance dimensions, enabling structured grading and calibration control.

Across all families, resource rules regulate whether examinees must rely on recalled doctrine or may apply provided authorities. This design feature prevents construct contamination by ensuring that tasks measure applied reasoning rather than resource navigation alone.

The detailed Blueprint is available at ncbex.org/sites/default/files/2025-07/NCBE-NextGen-UBE-Blueprint_5.pdf.

Validity and Inference Implications

The NextGen UBE is built as a layered measurement system:

Construct definitions specify what is measured.

The Blueprint translates those constructs into enforceable test specifications.

Item families operationalize those specifications into observable performance.

Together, these layers form a structured validity framework that links licensure purpose to score interpretation.

The inferential sequence proceeds as follows:

1. The licensure purpose requires assessment of minimum competence.
2. Practice analysis defines required knowledge and skill domains.
3. The Blueprint translates domains into enforceable structural specifications.
4. Item families operationalize those specifications into observable performances.
5. Standardized scoring and equating procedures support defensible score interpretations.

By constraining design at each level, the NextGen UBE reduces construct-irrelevant variance, strengthens cross-form comparability, and supports the interpretive claim that a passing score represents readiness for entry-level practice.

Section Composition and Timing Model

The NextGen UBE consists of three three-hour sections administered over one and a half days. Each section follows a fixed internal structure designed to balance breadth of doctrinal sampling with depth of applied skill measurement.

Within each three-hour section, examinees encounter a mix of question types:

- **40** standalone multiple-choice questions
- **2** integrated question sets
- **1** performance task

Across the full examination, this results in the following totals:

- **120** standalone multiple-choice questions
- **6** integrated question sets
- **3** performance tasks

The order of item types within each section is standardized. Questions may be answered in any order within the section, but all components must be completed within the fixed three-hour window. No carryover time is permitted between sections.

Time allocations were informed by pilot and prototype research. Based on observed performance data, examinees are expected, on average, to spend approximately

- **1.8** minutes per standalone multiple-choice question;
- **24** minutes per integrated question set; and
- **60** minutes per performance task.

These time expectations are not imposed constraints but empirically derived averages that informed section design. The goal is to ensure sufficient opportunity for skill demonstration while maintaining standardized time pressure consistent with licensure testing conditions.

Item-Assembly and Form-Construction Rules

Form construction follows predefined assembly constraints to preserve construct representation and comparability across administrations.

Standalone multiple-choice items are distributed across the eight Foundational Concepts and Principles in approximately equal proportions over the full examination. This distribution supports broad doctrinal sampling and provides the anchor necessary for stable equating.

Integrated question sets and performance tasks are assembled to ensure

- representation of multiple Foundational Skills;
- controlled variation in doctrinal context;
- balanced cognitive demand; and
- consistent scoring weight allocation.

Pretest items are embedded among the standalone multiple-choice questions and in one integrated question set per administration. Pretest items are indistinguishable from scored items and do not contribute to examinee scores. This embedded design supports continuous item calibration without altering examinee experience.

Performance tasks are pretested independently and incorporated into operational forms only after psychometric review.

All forms are constructed to meet predefined statistical and content specifications. Assembly rules constrain

- total score weight by item type;
- skill-domain representation;
- doctrinal coverage;
- response-format distribution; and
- partial-credit allocation structure.

These constraints ensure that each operational form represents the same underlying construct within defined tolerances.

Structural Controls and Comparability

The structural model supports three critical measurement goals:

Content Representativeness — ensuring adequate sampling of required domains

Score Stability — enabling reliable scaling and equating

Comparability Across Administrations — preserving interpretive consistency over time

The combination of fixed section composition, defined weightings, calibrated timing expectations, and constrained assembly rules provides the psychometric foundation necessary for defensible score interpretation.

This structural architecture also enables longitudinal comparability, supporting jurisdictions in making passing-standard and portability decisions with confidence that score meaning remains stable across administrations.

Part I Summary: Foundations and Design Controls

Part I of this report establishes the conceptual and structural foundation of the NextGen UBE.

The development process began with a defined licensure purpose: measurement of minimum competence for newly licensed lawyers. That purpose was translated into construct definitions grounded in empirical practice analysis and stakeholder input. The resulting domains of foundational knowledge and skills define the target of measurement.

The Blueprint operationalizes those domains into enforceable specifications governing content coverage, item-type weighting, and section composition. Item-family specifications further constrain how those specifications are translated into observable performance tasks. Together, these layered design controls establish a traceable chain from licensure purpose to exam structure.

By constraining construct definition, blueprint architecture, and form assembly rules, the examination design reduces construct-irrelevant variance and supports cross-form comparability. The foundational architecture described in Part I provides the structural basis for psychometric evaluation and score interpretation.

Part II describes how this architecture is implemented within a digitally integrated ecosystem.

Part II.
A Digital Evolution

A Digitally Native Ecosystem

The NextGen UBE is not a paper examination converted to digital format. It is a digitally native assessment ecosystem designed to support the full lifecycle of licensure administration—from candidate registration through grading and reporting—within an integrated, role-based architecture.

The system components referenced in this report are:

- **Jurisdiction Portal** – the administrative environment for jurisdiction administrators
- **Candidate Portal** – the candidate-facing readiness and status system
- **ITS Exam Day Portal** – the live monitoring interface used during administration
- **Delivery Platform** – the secure testing application through which content is presented and responses are captured
- **Grading Platform** – the digital scoring environment for constructed responses

Each component serves a distinct function. Together, they operate as an interconnected system through defined APIs and controlled data flows.

Jurisdiction Portal

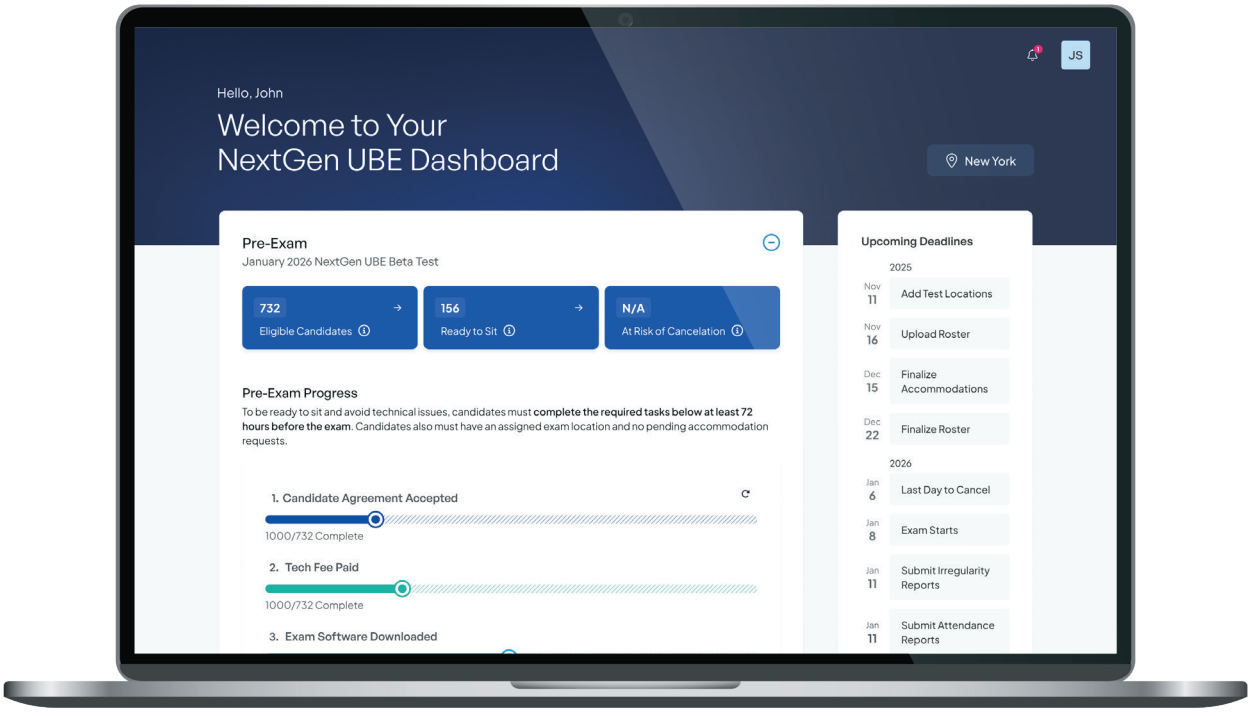
The Jurisdiction Portal functions as the centralized administrative environment for participating jurisdictions. It supports pre-exam preparation, live monitoring, and post-exam documentation within a unified system.

Historically, jurisdictions relied on distributed tools and manual reconciliation processes to manage candidate rosters, accommodations, attendance, and irregularities. The Jurisdiction Portal consolidates these processes into a single role-based interface, reducing fragmentation and improving visibility across the administration lifecycle.

Pre-Exam Administration

During the pre-exam phase, jurisdictions use the portal to upload and verify candidate rosters, assign testing locations, finalize accommodations determinations, and monitor completion of required readiness steps.

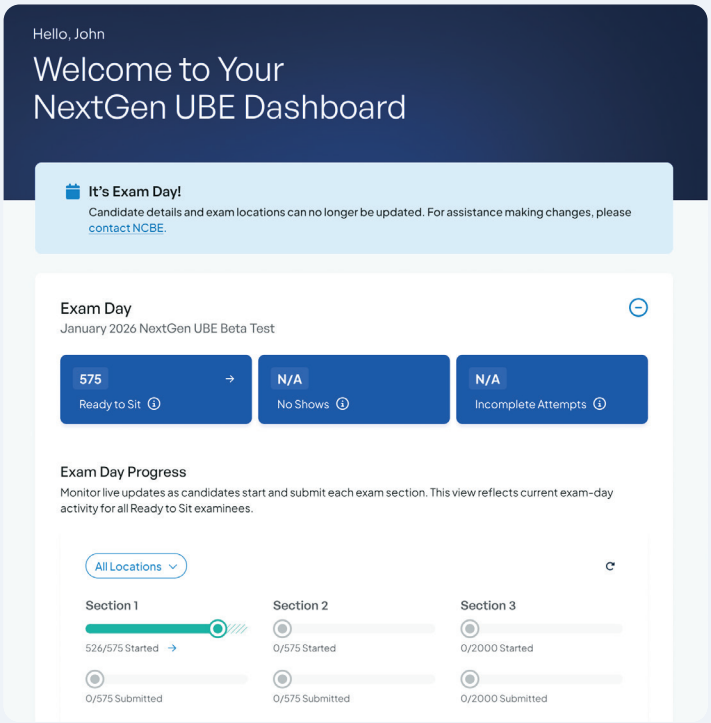
Administrative dashboards provide real-time visibility into eligibility confirmation and readiness status. These dashboards function as early warning systems, identifying candidates at risk of cancellation or incomplete preparation prior to exam day. Centralized deadline tracking reduces reliance on external spreadsheets or manual reconciliation.

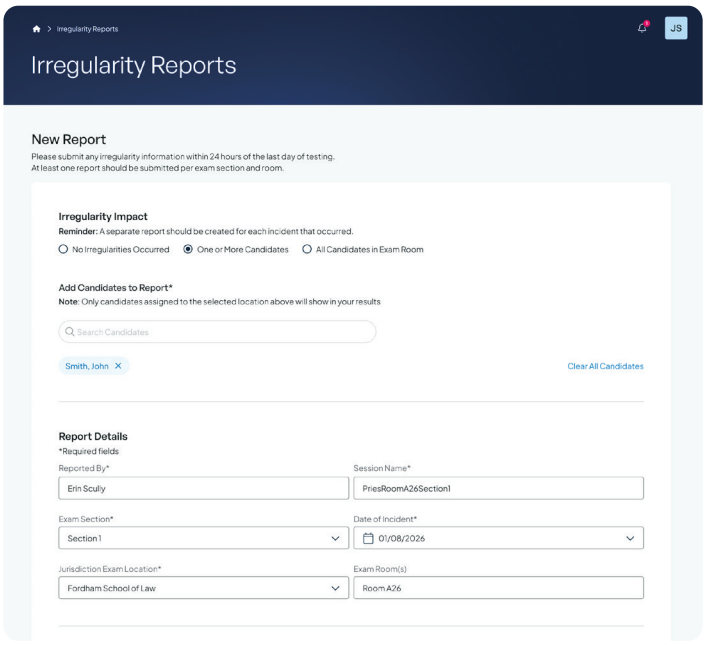


Exam-Day Monitoring

During exam delivery, the Jurisdiction Portal provides administrative visibility into examinee log-in status, attendance designation, section launch and submission status, and incomplete or interrupted sessions. Monitoring views may be filtered by testing location, allowing jurisdictions to oversee multiple sites simultaneously.

Integrated policy guidance and documentation tools support consistent implementation of administrative procedures during live testing.





Post-Exam Documentation

Following exam delivery, the portal supports attendance verification and structured irregularity reporting. Standardized reporting forms ensure consistent documentation of technical incidents, environmental disruptions, and administrative events.

All records are retained within the system, creating a complete administrative audit trail that supports reconciliation, review, and longitudinal operational analysis.

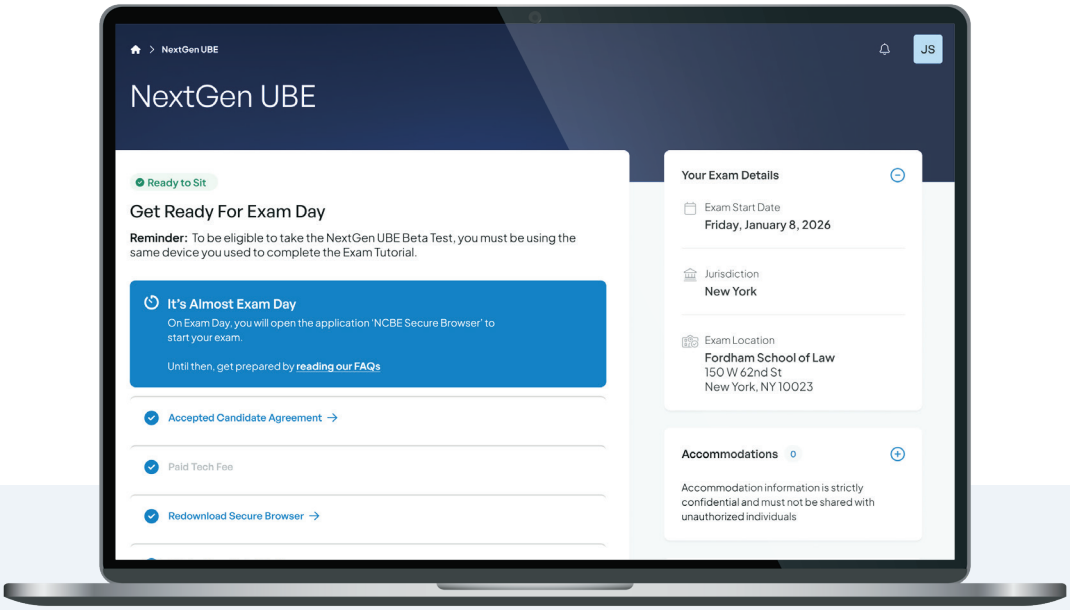
Across phases, the Jurisdiction Portal operates as the administrative command environment for the exam.

Candidate Portal

The Candidate Portal serves as the centralized candidate-facing system. It supports readiness preparation, accommodations visibility, logistics communication, and exam-day confirmation.

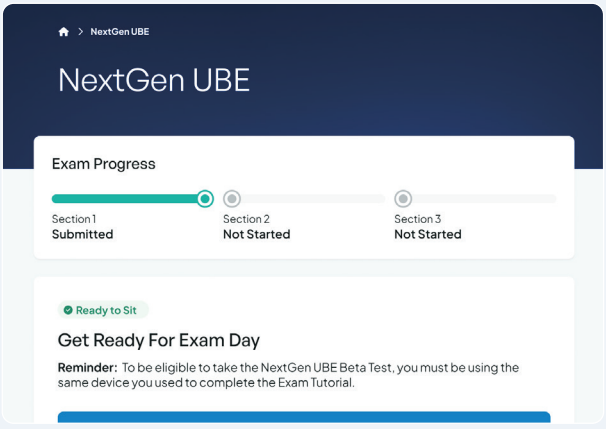
In prior administrations, readiness processes relied heavily on distributed communications and manual verification. The Candidate Portal consolidates required steps into a structured workflow with real-time status indicators.

Candidates must complete defined readiness steps—including acceptance of the Candidate Agreement, technology fee payment, secure browser installation, and completion of the exam tutorial—before being designated as meeting NCBE requirements. This designation reflects both jurisdiction eligibility confirmation and completion of NCBE readiness steps, ensuring alignment between candidate preparation and administrative approval. However, there may be additional steps a candidate must take with the jurisdiction before being confirmed as "Ready to Sit."



Accommodations determinations are visible within the portal. Automated validation processes flag duplicate registrations and inconsistent candidate records, supporting data integrity prior to administration.

On exam day, the portal provides confirmation of eligibility and post-section response-submission status without exposing exam content or responses. Candidates are directed to launch the secure delivery platform and may verify successful submission following each section.



By centralizing readiness and status visibility, the Candidate Portal reduces administrative ambiguity and improves structured preparation at scale.

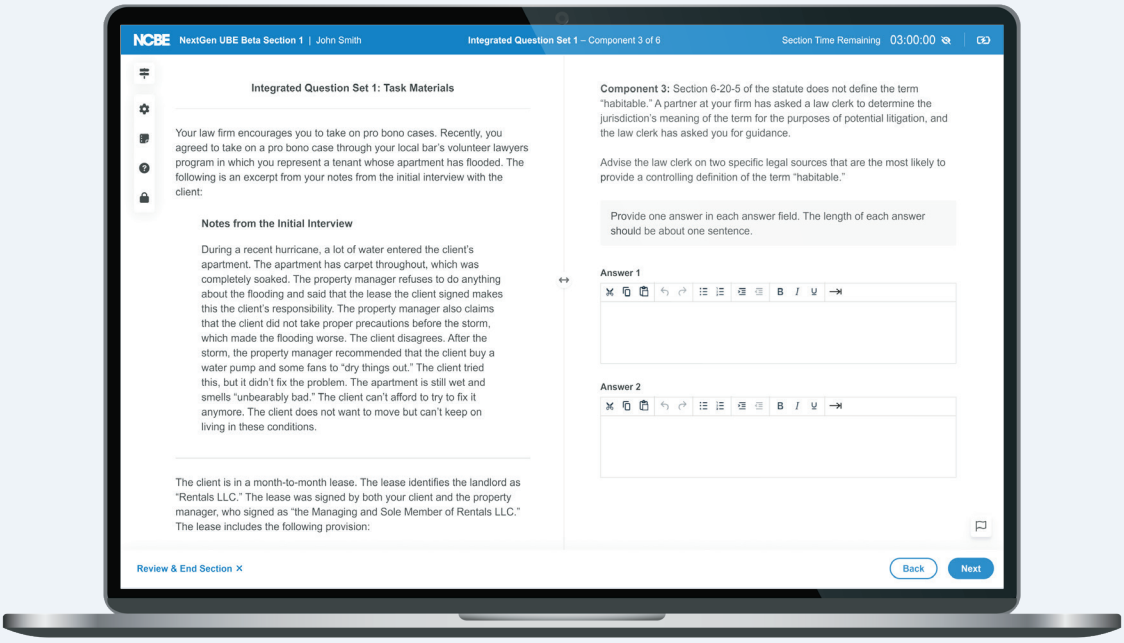
Delivery Platform

The delivery platform is the secure, candidate-facing testing engine. It presents all exam content and captures all responses within a controlled application environment.

It is architecturally distinct from the Jurisdiction Portal, Candidate Portal, and ITS Exam Day Portal. Those systems manage readiness and monitoring; the delivery platform executes the examination.

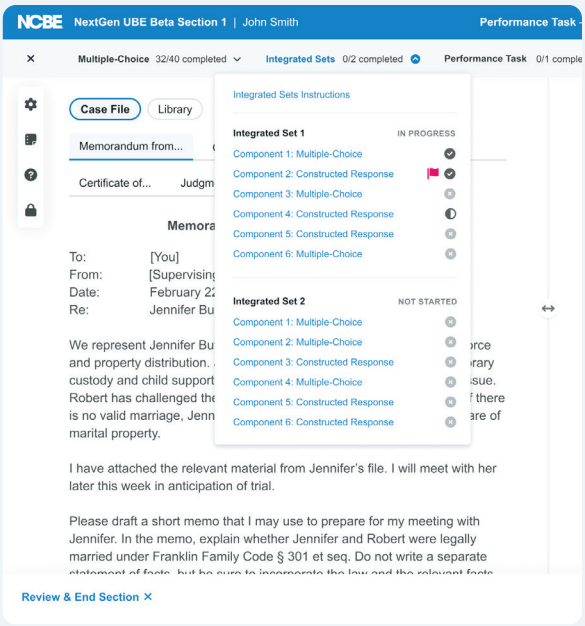
The platform supports all item types within a unified interface, including standalone selected-response questions, integrated question sets, standard performance tasks, and legal research performance tasks. Content is delivered within a secure browser environment designed to protect exam integrity and ensure consistent presentation.

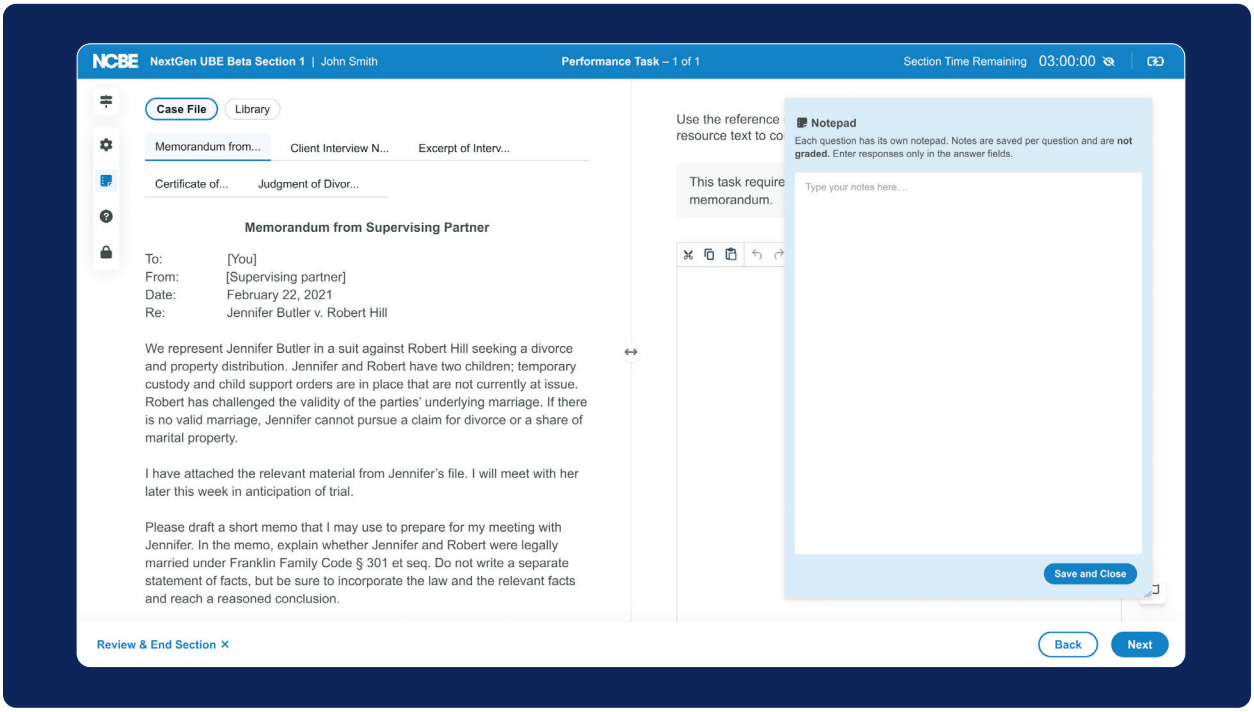
Examinees may navigate nonlinearly within sections. Integrated question sets and performance tasks use split-screen or tabbed layouts to present source materials alongside response fields. Structured review screens provide confirmation prior to submission.



Built-in tools—including highlighting, strike-through, notepad functionality, copy-paste, spell check, zoom controls, and timer visibility—are embedded within the application. Accessibility features such as adjustable contrast and text resizing are available to all examinees.

Accommodations configurations—including extended time, stop-the-clock breaks, speech-to-text, and text-to-speech—are implemented through platform settings rather than separate systems. This unified configuration model preserves construct consistency across standard and accommodated administrations.





Responses are captured continuously during active testing and submitted through a structured workflow that includes confirmation screens and upload verification indicators.

ITS Exam Day Portal

The ITS Exam Day Portal functions as the real-time monitoring interface used during administration. It provides operational oversight of testing activity across locations.

Administrators can monitor log-in status, attendance, section progress, break status, and submission completion. Filters and dashboards provide summary metrics while allowing drill-down into individual candidate status.

The ITS Exam Day Portal does not expose exam content. Its purpose is operational monitoring, not content interaction.

Grading Platform

The grading platform is the secure scoring environment used for constructed responses. It manages assignment, independent double grading, reconciliation workflows, and oversight monitoring.

Following reconciliation of delivery data, responses are released into the grading platform. The system assigns responses according to defined scoring models, preserving independence of initial evaluations. Analytic rubrics and performance descriptors are presented within the interface.

The platform automatically identifies score discrepancies exceeding defined thresholds and routes responses through structured reconciliation workflows. Administrative dashboards allow authorized personnel to monitor completion rates, reconciliation volume, and timeline adherence.

All grading actions are logged, preserving a complete audit trail of scoring decisions.

API Integration and Data Flows

The ecosystem components are integrated through defined application programming interfaces (APIs) and controlled data exchanges. Each system maintains role-based boundaries while sharing necessary data elements to support the examination lifecycle.

Key integrations include

- transmission of eligibility and readiness data from the Candidate Portal to the Jurisdiction Portal;
- secure launch authorization from portals to the delivery platform;
- real-time status signals from the delivery platform to the ITS Exam Day Portal;
- structured response transfer from the delivery platform to the grading platform; and
- reconciliation and scoring data returned to administrative systems.

These APIs enforce validation rules at each transfer point. Data exchanges are structured, logged, and subject to security controls. No system operates as an isolated environment, yet no system exposes data beyond its defined role permissions.

This architecture supports

- administrative visibility without content exposure;
- separation of candidate interaction from scoring evaluation;
- secure longitudinal record management; and
- reduced reliance on manual reconciliation.

The digital ecosystem therefore functions as an integrated, lifecycle-based infrastructure rather than a collection of discrete tools.

Security, Privacy, and Compliance

Security and data integrity are embedded within the system architecture. Access across all components is role based. Sensitive candidate information is restricted according to administrative function. Content exposure is limited to authorized roles within defined time windows.

Structured logging across systems preserves auditability. Data governance controls regulate storage, transfer, and retention of candidate and scoring records.

Accessibility and Accommodations

Accessibility is embedded at two levels:

1. Universal design features available to all examinees
2. Configurable accommodations settings delivered within the same platform environment

By implementing accommodations through configuration rather than alternative systems, the ecosystem preserves equivalence of construct measurement while supporting individualized access needs.

Operational Administration

The operational administration of the NextGen UBE is designed as a structured, observable, and auditable process. Digital delivery does not reduce the need for administrative controls; it requires more explicit ones. The operational model integrates jurisdiction oversight, centralized monitoring, standardized policies, and structured documentation to preserve exam integrity across multiple sites.

Administration is organized across three phases: pre-exam preparation, live delivery oversight, and post-exam reconciliation.

Pre-Exam Processes and Controls

Pre-exam administration is governed by defined readiness requirements, structured eligibility verification, and coordinated jurisdiction oversight.

Jurisdictions upload and verify candidate rosters through the Jurisdiction Portal. Eligibility to sit is confirmed through role-based validation, ensuring that candidates are both administratively approved and technically prepared. Readiness status reflects completion of required steps within the Candidate Portal and confirmation by the jurisdiction.

Accommodations determinations are finalized prior to administration. Approved accommodations are configured within the delivery platform rather than through alternative systems, preserving measurement equivalence while ensuring individualized access.

Technical readiness is confirmed through secure browser installation and completion of the exam tutorial. These steps function both as system validation and candidate familiarization mechanisms.

Pre-exam dashboards provide visibility into incomplete readiness requirements, enabling intervention before exam day. This reduces the likelihood of last-minute administrative disruption and supports uniform preparation across jurisdictions.

Exam-Day Administration and Proctoring

Exam-day administration operates through coordinated local proctoring and centralized digital monitoring.

Local proctors oversee physical testing environments in accordance with standardized administration manuals. Their responsibilities include identity verification, enforcement of testing rules, management of start and stop times, and documentation of irregularities.

Simultaneously, the ITS Exam Day Portal provides real-time administrative visibility into testing activity across locations. Administrators can observe log-in status, attendance designation, section progression, break status, and submission confirmation without access to exam content.

This dual-layer oversight model preserves

- local control of physical testing conditions;
- centralized visibility into digital session status; and
- separation of operational monitoring from content access.

Stop-the-clock breaks and other time-based accommodations are monitored through system indicators, ensuring that extended-time calculations are implemented accurately and consistently.

Structured confirmation workflows at the end of each section ensure that examinees intentionally submit responses. Submission status is visible to both candidates and administrators, reducing ambiguity regarding response capture.

Incident Classification and Escalation

Irregularities are classified using structured categories within the Jurisdiction Portal. Incidents may include technical interruptions, environmental disruptions, administrative errors, or examinee conduct issues.

Documentation is completed using standardized digital forms, which capture

- the nature of the incident;
- timing relative to section progress;
- resolution steps taken; and
- impact on candidate testing status.

Defined escalation pathways govern when incidents require jurisdiction-level review, NCBE oversight, or technical vendor intervention. This structured classification model promotes consistency across jurisdictions and supports post-exam review.

Because the delivery platform continuously captures response data during active testing, most technical interruptions do not result in data loss. Session recovery protocols allow candidates to resume testing under controlled conditions, subject to administrative authorization.

Monitoring and Transparency

Operational transparency is embedded into the administration model. Administrators have real-time visibility into attendance, session progress, and submission status. Candidates receive confirmation of response submission at the conclusion of each section.

Audit logs preserve

- log-in attempts;
- session launches;
- break initiations and returns;
- submission time stamps; and
- administrative status changes.

These logs support post-administration reconciliation and, when necessary, investigative review.

Post-Exam Processes and Controls

Following completion of testing, jurisdictions verify attendance records and finalize irregularity documentation within the Jurisdiction Portal. These records form the official administrative record for each administration.

Response data are reconciled and validated prior to release into the grading platform. This reconciliation ensures that all submitted responses are accounted for and that incomplete or interrupted sessions are reviewed in accordance with established policies.

The separation between delivery and grading environments preserves scoring independence. Administrative oversight concludes once response reconciliation is complete and grading workflows are initiated.

Operational Integrity

The operational model is designed to achieve four objectives:

1. Preserve exam security

2. Ensure consistent policy implementation

3. Maintain accurate and complete response capture

4. Support auditability and post-administration review

The beta administration confirmed that these controls function as intended under live, multi-site conditions. Jurisdictions were able to monitor readiness, oversee delivery, document incidents, and reconcile records within a unified digital framework.

The operational administration structure supports scalable, consistent delivery while maintaining the evidentiary controls required for defensible licensure testing.

Scoring and Grading Platform

Scoring Architecture

The NextGen UBE employs a hybrid scoring model combining machine scoring for selected-response items and analytic human scoring for constructed-response components. This model reflects deliberate design decisions balancing measurement breadth, construct representation, and reliability.

Standalone multiple-choice questions are scored centrally using defined answer keys and automated scoring algorithms. Select-two items permit partial credit in accordance with predefined scoring rules. These items contribute 49% of the total exam score and provide the primary anchor for equating and scale stability.

Integrated question sets and performance tasks are scored using structured analytic rubrics within a centralized digital grading platform. Constructed-response components collectively account for 51% of the total score, with integrated question sets contributing 21% and performance tasks contributing 30%.

All scores are ultimately reported as a single scaled score within the defined reporting range.

Grading Workflow

Constructed responses are released into the grading platform only after reconciliation of delivery data and confirmation of complete response capture. Responses are anonymized prior to scoring. Graders do not have access to candidate-identifying information.

Each constructed response is independently scored by two trained graders. Independent double grading is the default scoring model. Graders enter scores directly into the digital interface, which presents

- the examinee response;
- the applicable analytic rubric;
- performance descriptors; and
- scoring guidance materials.

The interface does not expose other graders' scores during initial evaluation. This preserves independence of judgment and reduces anchoring effects.

Reconciliation and Adjudication

Following initial scoring, the system automatically compares the two assigned scores for each response. Predefined discrepancy thresholds determine whether reconciliation is required.

When score differences fall within tolerance, scores are combined according to established aggregation rules. When discrepancies exceed defined thresholds, responses are routed through structured reconciliation workflows. These may involve

- consensus review by graders;
- review by a designated team leader; or
- adjudication by a senior scoring authority.

All reconciliation decisions are recorded within the system. The platform preserves both initial scores and final resolved scores, maintaining a complete audit trail.

This structured reconciliation model serves two functions: it improves inter-rater reliability (consistency of grader determinations) and provides documented evidence of oversight in cases of score variance.

Rubric Design and Analytic Scoring

Constructed-response items are evaluated using analytic rubrics developed during content creation. Rubrics allocate discrete point values to defined performance elements. Partial credit is awarded in full-point increments according to rubric criteria.

Analytic scoring supports

- transparent linkage between performance and score;
- consistent grader calibration;
- structured feedback during training; and
- improved score reliability relative to holistic judgment models.

Rubrics are accompanied by representative performance examples and scoring commentary used during grader training and calibration.

Oversight and Monitoring

The grading platform provides real-time oversight dashboards available to authorized supervisory personnel. These dashboards display

- completion rates by item and team;
- distributional patterns across graders;
- reconciliation volume; and
- timeline adherence.

Supervisory monitoring supports identification of anomalous scoring patterns and enables intervention where necessary. However, oversight visibility does not interfere with independence of initial scoring decisions.

Data Integrity and Security

Access to the grading platform is role-based and restricted to authorized personnel. Responses and scoring records are maintained within a secure environment governed by established data governance protocols.

All scoring actions—including initial scores, reconciliations, and administrative interventions—are logged. Audit logs preserve time stamps, user identifiers, and scoring events, ensuring traceability.

Constructed-response data remain separate from delivery systems during grading. This architectural separation preserves scoring independence and reduces risk of unauthorized content exposure.

Measurement Implications

The scoring architecture supports the scoring and generalization inferences underlying the validity argument. Independent double grading, structured discrepancy thresholds, analytic rubrics, and centralized oversight collectively strengthen reliability and consistency of constructed-response scores.

The integration of machine-scored selected-response items with analytically scored constructed responses allows the exam to achieve both domain breadth and performance depth while maintaining psychometric control.

Grader Support

Grader Qualification and Selection

Constructed-response grading is performed by jurisdiction-appointed graders who meet defined eligibility criteria established by participating jurisdictions. All graders must complete standardized training and qualification requirements prior to scoring operational responses.

Eligibility alone does not authorize scoring. Access to operational scoring environments is contingent upon successful completion of training and calibration benchmarks described below.

Training Architecture

Grader training is structured and item-specific. Training materials are developed during the content finalization process and include

- construct definition and skill targets;
- analytic rubric structure;
- performance descriptors for each score point;
- annotated exemplar responses across performance levels; and
- common error patterns and borderline case guidance.

Training is delivered through a centralized digital module. Graders must demonstrate understanding of rubric criteria and scoring standards before proceeding to calibration exercises.

The objective of training is not familiarity with content alone, but alignment with performance-level distinctions defined in the scoring model.

Calibration and Qualification

Following training, graders participate in calibration rounds using pre-scored benchmark responses. These benchmark responses represent defined score points across the rubric scale. This structured qualification process supports the scoring inference by reducing variance attributable to rater interpretation.

Ongoing Monitoring and Drift Control

Grading quality is monitored continuously during operational scoring.

The grading platform supports

- real-time tracking of score distributions by grader;
- monitoring of reconciliation rates;
- detection of anomalous scoring patterns; and
- review of outlier scoring behaviors.

Supervisory personnel review scoring metrics at defined intervals. When patterns suggest possible scoring drift, targeted recalibration may be initiated.

Drift controls may include

- review of scoring guides;
- review of assessment responses; or
- discussion with team or team leaders.

These interventions are structured and documented within the system.

Reconciliation Oversight

Independent double grading inherently produces reconciliation cases. Reconciliation volume and resolution patterns are monitored to identify potential systemic variance.

Team leaders and adjudicators operate under defined review protocols. Reconciliation decisions must align with rubric definitions and are subject to audit review.

The system preserves

- initial independent scores;
- final reconciled scores;
- identity of scoring personnel; and
- time-stamped scoring actions.

This audit trail supports defensibility and transparency.

Reliability and Measurement Implications

The grader support architecture strengthens the generalization inference underlying constructed-response scoring.

By combining standardized training, structured calibration benchmarks, qualification thresholds, continuous distribution monitoring, and controlled reconciliation workflows, the system reduces variability attributable to rater inconsistency and supports defensible reliability estimates.

Grader support is therefore not an auxiliary process but an integral component of the measurement system.

Part II Summary: Operational and Measurement Infrastructure

Part II describes the digital, administrative, and scoring infrastructure through which the NextGen UBE is delivered and evaluated.

The ecosystem integrates candidate readiness, jurisdiction oversight, secure content delivery, structured monitoring, and analytic scoring within defined system boundaries. APIs connect system components while preserving role-based access controls and content security. Administrative and scoring environments are intentionally separated to protect independence of evaluation.

Operational administration is governed by standardized policies, structured monitoring tools, and documented incident workflows. Response-capture and reconciliation processes ensure completeness and auditability of examinee data.

The scoring architecture combines automated selected-response scoring with independent double grading of constructed responses, supported by analytic rubrics, qualification thresholds, reconciliation controls, and continuous oversight monitoring. Audit trails are preserved at each stage.

Collectively, these systems provide the infrastructure necessary to support the scoring, generalization, and comparability inferences underlying the exam's validity argument.

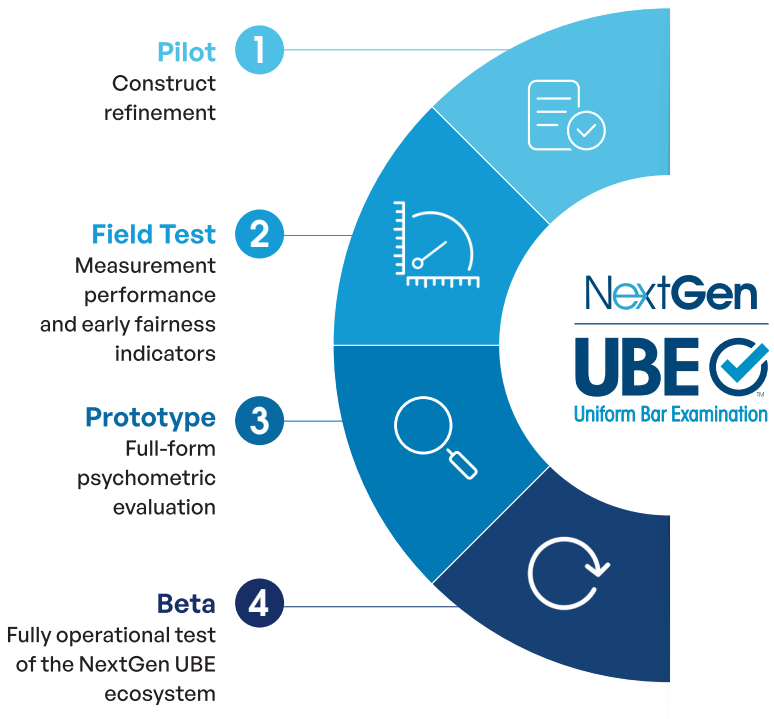
Part III presents the empirical evidence generated through the testing arc to evaluate how these design and infrastructure elements perform under live conditions.

Part III. Testing and Development

Testing Arc and Empirical Evidence

The NextGen UBE was developed through a staged empirical program designed to evaluate construct representation, scoring reliability, structural coherence, and operational performance under progressively realistic conditions. Each phase of the testing arc increased in scale, integration, and psychometric scrutiny.

The testing arc comprised four sequential phases: pilot testing, field test, prototype exam, and beta administration. Each phase generated distinct forms of evidence and informed subsequent design decisions.



Pilot Phase: Construct Refinement

The pilot phase evaluated early item families and task structures under controlled conditions. The objective was to assess whether tasks elicited the intended knowledge–skill integration and whether scoring rubrics differentiated performance levels as designed.

Analyses included item difficulty and discrimination statistics, item-total correlations, preliminary dimensionality assessment (i.e., assessing whether a single or multiple constructs were being measured), and rubric-level score functioning. Constructed-response pilots examined whether analytic scoring criteria produced stable score distinctions aligned with defined performance descriptors.

Pilot findings led to revisions in prompt clarity, rubric specificity, timing assumptions, and task instructions. The emphasis at this stage was design precision rather than large-scale stability.

Field Test Phase: Measurement Performance and Early Fairness Indicators

The January 2024 field test administered five two-hour forms to more than 4,000 final-year law students and recent graduates across 88 law schools. The forms included 155 questions spanning standalone selected-response items, drafting sets, counseling sets, standard performance tasks, and a legal research performance task.

Item Difficulty and Distribution

When evaluating item difficulty, mean p-values (representing the proportion of examinees who answer an item correctly) ranged from 0.48 for single-select selected-response items to 0.65 for multiple-select items. P-values for constructed-response formats ranged between 0.51 and 0.63. These values fall within conventional target ranges for high-stakes licensure assessments, indicating appropriate calibration of difficulty.

Items previously used operationally on the MBE demonstrated lower p-values in the field test sample than in live administrations. Follow-up analyses of response time and omission patterns did not indicate disengagement; rather, the difference was consistent with lower preparation levels among volunteer participants.

Overall, the distribution of item difficulty supported continued development without evidence of systematic miscalibration.

Response Time Analysis

Mean observed response times were as follows:

1.3 minutes per standalone selected-response item	17.6 minutes per drafting set	15.1 minutes per counseling set	42.5 minutes per performance task

Ninetieth-percentile response times modestly exceeded initial design assumptions for selected-response items and drafting sets but were below target timing for counseling sets and performance tasks. These findings informed refinements to section timing models in subsequent prototype development.

No evidence of systemic speededness (i.e., time limits affecting examinee performance) was observed.

Grading Model Evaluation

The field test evaluated the transition from a relative grading model to an absolute analytic grading model for constructed responses.

Approximately 37,000 responses were scored by 61 graders from 27 jurisdictions. Forty-four percent of responses were independently double graded, and approximately 30% of those required adjudication.

Variance analyses demonstrated that the majority of score variance was attributable to examinee performance rather than rater effects. Grader feedback indicated that analytic rubrics increased clarity and alignment with performance criteria, although initial familiarization with the rubrics required additional cognitive effort.

Timing analyses suggested that, after orientation, total grader hours under the double-grading model would approximate legacy single-grading workloads.

Subgroup Performance

Standardized mean differences were examined across Foundational Skills and demographic variables. Among recent graduates—the group most analogous to bar examinees—observed effect sizes were generally small. Differences by race/ethnicity ranged from 0.15 to 0.44, remaining within the small-effect range typically observed in licensure testing.

These analyses provided preliminary evidence that the new item formats did not materially amplify subgroup performance disparities.

Field Test Implications

The field test demonstrated appropriate difficulty calibration, feasible timing, workable analytic grading, and no evidence of material amplification of subgroup performance gaps. These findings supported progression to full-form prototype evaluation.

Full statistical documentation of the field test analyses is at nextgenbarexam.ncbex.org/reports/nextgen-research-brief-field-test/.

Prototype Phase: Full-Form Psychometric Evaluation

The October 2024 prototype exam represented the first administration of a complete nine-hour NextGen form under operationally representative conditions. The analysis sample included more than 2,000 participants drawn from July 2024 bar examinees, with ability and demographic distributions aligned to historical operational populations.

The prototype phase served five primary purposes:

1. Evaluate psychometric performance of a complete form
2. Establish the NextGen reporting scale
3. Conduct concordance analyses with the legacy UBE
4. Inform standard-setting studies
5. Evaluate operational feasibility under scaled conditions

Item Difficulty and Calibration

Among 129 scored items, mean p-values ranged from 0.55 for short-answer constructed responses to 0.73 for multiple-select selected-response items. Ninety-two percent of items fell within predefined selection ranges used for item flagging and review.

Item calibration employed Rasch modeling for dichotomous items (single-select selected-response questions) and partial-credit modeling for polytomous items (multiple-select selected-response questions and constructed responses). All calibrated items demonstrated acceptable fit statistics within established thresholds for high-stakes assessment. Difficulty parameters ranged approximately from −2.0 to +2.0 logits, indicating appropriate spread across the ability continuum.

Short-answer constructed-response items exhibited slightly higher average difficulty parameters than selected-response formats, consistent with construct expectations.

Item Discrimination

Item-total correlations were generally acceptable across formats. Extended constructed-response items demonstrated higher mean discrimination values, reflecting their broader score-point scales and larger contribution to total score variance.

Only isolated items fell below conventional discrimination thresholds and were flagged for content review.

Dimensionality and Structural Coherence

Exploratory and confirmatory factor analyses were conducted to evaluate the structural coherence of the integrated exam. Results supported interpretation of the examination as measuring a unified competence construct with correlated facets corresponding to item families.

No evidence emerged of problematic local item dependence (i.e., responses to some items being influenced by other items) beyond what would be expected given shared stimulus materials within integrated sets.

Reliability

Section-level and total-test reliability coefficients met expectations for high-stakes licensure testing. Reliability estimates supported stable score interpretation at decision thresholds relevant to pass/fail determinations.

Differential Item Functioning

Differential item functioning (DIF) analyses were conducted across gender and other demographic groups. Isolated items demonstrated statistically significant DIF and were flagged for sensitivity review; however, no systemic pattern of DIF was observed across item types.

Response Time and Speededness

Observed 90th-percentile response times exceeded design targets for certain performance tasks, with some extended tasks approaching 74 minutes against a 60-minute design expectation. These findings informed adjustments to timing models and administration guidance prior to beta.

Prototype Implications

The prototype administration confirmed appropriate item calibration, acceptable discrimination, coherent dimensional structure, stable reliability, manageable DIF patterns, and operational feasibility at full-form scale. The examination performed within psychometric tolerances expected for high-stakes licensure assessment.

The prototype administration resulted in several key outputs:

- establishment of the NextGen reporting scale (500–750)
- full Rasch calibration of scored items
- dimensionality and local item dependence analysis
- reliability estimation under full-length conditions
- national standard-setting study
- concordance analysis with the legacy UBE

Using hybrid Rasch modeling, item parameters and examinee abilities were estimated on a common scale. Reliability estimates surpassed typical thresholds for high-stakes licensure exams. Dimensionality analyses supported interpretation of the exam as measuring a unified construct of minimum competence for entry-level legal practice, while appropriately accounting for item set structure.

Following the prototype exam, a national standard-setting study was conducted using modified Angoff and Hofstee methods. Panelists evaluated the performance of a minimally qualified candidate. Outcome modeling across jurisdictions supported a recommended passing score range of 610–620.

Although the prototype exam provided robust psychometric evidence, it was administered prior to final blueprint stabilization and before full operational maturity of the digital ecosystem. For that reason, the January 2026 beta administration was designed to confirm these findings under launch-ready conditions. The next section provides a detailed analysis of the beta administration of the NextGen UBE.

Beta Administration: Operational Confirmation and Recommended Passing Score Range Validation

The January 2026 beta administration represented the final large-scale, fully integrated test of the NextGen UBE ecosystem prior to operational launch. Unlike prior stages of testing, which focused on specific research objectives (e.g., item functionality, scale establishment, or standard setting), the beta was designed to evaluate the complete system end to end: content, delivery platform, scoring workflows, jurisdictional interfaces, and data capture.

Administration Sites and Jurisdiction Participation

The beta was administered across five physical testing locations in four jurisdictions: Texas, Florida, New York, and two sites in Massachusetts.

All sites adhered to standardized proctoring protocols. Site directors were trained on incident response, candidate identification procedures, accommodations management, and technology contingency workflows. Jurisdiction representatives observed the administration in real time, providing an additional layer of operational evaluation.

The multi-site structure was intentional. It ensured that psychometric findings were not limited to a single geographic context and allowed for observation of environmental variability, including differing site infrastructure, staffing models, and logistical demands.

Administration Timeline and Structure

1,512 examinees began the administration. All examinees had previously sat for the July 2025 administration of the legacy bar exam. While the majority of examinees tested under standard conditions, examinees who had been approved for accommodations in the July administration were

granted them in the beta exam. Standard delivery of the beta examination was administered over two days, consistent with the operational design of the NextGen UBE. Some examinees approved for an extended-time accommodation began the assessment a day early, bringing their total testing time to three days.

Time allocations were identical to those anticipated for operational administration. Break structures, log-in procedures, candidate identification checks, and session transitions were all implemented using the production platform and operational workflows.

This fidelity to operational procedures was critical. Psychometric results obtained under idealized or laboratory conditions do not necessarily generalize to real-world testing contexts. By administering the beta under operationally realistic conditions, NCBE ensured that measurement properties were evaluated within the same constraints that will exist at launch.

Accommodations

Examinees approved for testing accommodations received extended time or other approved modifications consistent with jurisdictional policies. Accommodations workflows were processed through the Candidate Portal and tracked through jurisdiction dashboards.

From a psychometric perspective, the accommodations procedures were important for two reasons:

1. They allowed evaluation of the platform's ability to manage individualized timing conditions without compromising data integrity.
2. They ensured that reliability and item-functioning analyses included examinees tested under common accommodations conditions, thereby reflecting real-world score interpretation contexts.

Accommodations implementation proceeded without disruption. Timing controls, session extensions, and response-capture mechanisms functioned as intended.

Technology Stability and Failover Testing

A distinctive feature of the beta administration was the intentional testing of failover procedures. In selected sessions, controlled interruptions were introduced to simulate

- temporary loss of internet connectivity;
- power interruption; and
- device transitions.

The delivery platform's response-preservation mechanisms were activated, and examinees resumed testing without loss of data.

All intentionally introduced failover events were successfully mitigated. No response data were lost. From a measurement standpoint, this stability is critical: interruptions that result in response loss can distort item difficulty, affect response-time distributions, and compromise score comparability.

The beta demonstrated that the technical delivery system supports uninterrupted measurement under adverse but realistic conditions.

Participation and Completion

A total of 1,512 examinees began testing. Of those, 1,500 completed all exam sections, yielding a completion rate of 99.2%. The 12 examinees who did not complete the exam either did not attend some sections or showed up late and were not permitted to sit. In no cases was the lack of completion due to technology or assessment failures.

High completion rates reduce the risk of attrition bias. When substantial numbers of examinees withdraw or fail to complete sections, score distributions can become distorted, artificially affecting reliability estimates and outcome modeling.

The near-complete participation observed in the beta strengthens confidence in the stability of psychometric estimates and reduces concerns regarding non-random missing data.

Sample Characteristics

The beta sample was intentionally stratified across key demographic and academic characteristics to ensure balanced representation across forms. Stratified assignment reduces the likelihood that observed form-level differences are attributable to subgroup distribution rather than item functioning.

The size of the beta sample for Form 1 was sufficient for stable Rasch calibration, reliable item-parameter estimation, DIF analysis, and outcome modeling across cut scores. For Forms 2 and 3, sample sizes primarily supported pretesting and more limited analyses, which was the intention when recruiting given that the items on these forms were primarily pretest items.

Relative to the prototype sample, the beta administration reflected differences in examinee readiness and familiarity with the NextGen structure. This shift is relevant in interpreting performance distributions and response-time stability.

Complete demographic breakdowns are presented in the NextGen Beta Test: Report on End-to-End Ecosystem Performance (available at ncbex.org/sites/default/files/2026-03/NCBE-NextGen-UBE-Beta-Report.pdf). For purposes of this technical report, it is sufficient to note that representation was adequate to support subgroup and fairness analyses.

Examination Forms and Distribution

The January 2026 beta included three complete NextGen UBE forms.

Form Structure

Each form adhered to the finalized NextGen content blueprint and included 120 multiple-choice questions, six integrated question sets, and three performance tasks. Forms were separated into three sections, with each section including 40 multiple-choice questions, two integrated question sets, and one performance task.

Form 1 was designated as the primary operational form, and included both operational and pretest content. Forms 2 and 3 mirrored the structural architecture of Form 1 and included additional pretest content to support continued item-pool expansion.

Form 1 was assembled to match the NextGen blueprint specifications, but Forms 2 and 3 deviated from the blueprint in terms of content representation in order to target content pretesting needs.

Examinee Distribution Across Forms

Examinees were assigned using a stratified allocation model to balance subgroup characteristics across forms. Balanced assignment is essential for accurate cross-form psychometric comparisons, detection of item-performance differences, and fairness analyses.

Table 1: Form Distribution

Form	Number Completed All Sections	% of Total
Form 1	1,038	69.2%
Form 2	232	15.5%
Form 3	230	15.3%

Completion rates were consistent across forms and ranged from 99% to 100%, indicating no form-specific operational disruptions.

Psychometric Properties of the January 2026 Beta Administration with Comparative Analysis to the October 2024 Prototype Exam

Introduction to the Psychometric Comparison

The October 2024 prototype administration established the NextGen reporting scale, calibrated items under the Rasch model, estimated reliability, and supported national standard setting. The January 2026 beta administration was designed to replicate these analyses under fully operationally representative conditions.

Replication is essential to sound measurement practice. A single administration, even one conducted carefully, does not provide enough evidence of long-term stability. The beta administration acted as a confirmation phase, evaluating whether the psychometric properties observed during the prototype exam continued when

- the content blueprint was finalized;
- the digital platform was fully optimized;
- scoring processes were scaled;
- examinee familiarity and preparedness potentially varied; and
- pretest integration expanded the item pool.

Throughout this section, results from the beta administration are presented alongside corresponding prototype metrics. Based on the maturation of the ecosystem, NCBE's expectation is that performance will be relatively stable between the legacy exam and NextGen.

Item and Item Set Performance

Item Difficulty (Classical Statistics)

In classical statistics, item difficulty is typically expressed as a p-value, representing the proportion of examinees who answer an item correctly. For example, a p-value of 0.70 indicates that 70% of examinees answered correctly. In this context, a distribution of difficulty values is desirable; items should neither be uniformly easy nor uniformly difficult.

The table below presents mean classical difficulty statistics for selected-response items.

Table 2: Classical Item Difficulty (prototype vs. beta)

Summary Statistics of Operational Item P-Value Statistics by Item Type (prototype vs. beta Form 1)

Item Type	Prototype			Beta Form 1		
	Mean	Min.	Max.	Mean	Min.	Max.
Constructed Response: Medium Length	0.62	0.58	0.66	0.73	0.68	0.77
Constructed Response: Short Answer	0.55	0.29	0.72	0.56	0.28	0.9
Constructed Response: Extended Length	0.61	0.49	0.67	0.66	0.57	0.71
Selected Response: Multi Select	0.73	0.59	0.93	0.71	0.54	0.91
Selected Response: Single Select	0.60	0.31	0.88	0.60	0.21	0.88

The beta administration demonstrated stable item-difficulty distributions that replicated prototype findings under operational conditions. Distributions remained centered within expected ranges, with the majority of items falling between 0.30 and 0.90, a range that limits extreme items and supports greater measurement information.

Observed differences fell within expected variation across administrations and did not indicate any systematic shift in difficulty. Such variation is expected across administrations and may reflect differences in samples and forms.

Together, these findings confirm that the reporting scale remains appropriately targeted to the minimally qualified candidate.

Rasch Difficulty Parameters and Model Fit

While classical p-values provide intuitive indicators of difficulty, Rasch modeling offers a more robust framework for scale calibration. Under the Rasch model, item difficulty and examinee ability are estimated on a common logit scale.

A logit represents a unit of measurement on the latent ability continuum. Items with higher logit values are more difficult; those with lower values are easier. Importantly, Rasch calibration supports invariance, meaning item difficulty estimates remain stable across samples when model fit assumptions are satisfied.

Model fit is evaluated using infit and outfit mean square statistics. Values near 1.0 indicate expected response behavior. Values substantially above 1.0 may indicate unpredictability, while values substantially below 1.0 may indicate redundancy.

Table 3: Rasch Item Difficulty and Model Fit (prototype vs. beta)

Metric	Prototype	Beta
Mean Logit Difficulty	0.00	-0.17
% Flagged for Misfit	0%	0%

Rasch difficulty estimates from the beta administration aligned closely with prototype calibration, confirming stability of the measurement scale under operational conditions. No items were flagged for misfit in either administration, indicating that response patterns were consistent with model expectations.

This stability supports two critical conclusions:

1. The underlying measurement model continues to describe examinee response behavior appropriately.
2. The NextGen reporting scale remains structurally stable under operational conditions.

Item Discrimination and Item-Total Correlations

Item discrimination reflects the degree to which performance on an item aligns with overall test performance. One commonly used metric is the item-total correlation.

An item-total correlation measures the strength of association between responses to a single item and total test score. Higher values indicate that an item differentiates effectively between higher- and lower-performing examinees.

Table 4: Item-Total Correlations (prototype vs. beta)

Item Type	Prototype Mean r	Beta Mean r	Δ
Selected Response: Single Select	0.25	0.24	-0.01
Selected Response: Multi Select	0.30	0.29	-0.01

Item discrimination remained stable across administrations, with beta results closely matching prototype performance and only minor differences across item types. Such small variations are expected across administrations.

Importantly, no degradation in discrimination was observed, indicating that measurement precision has been preserved following blueprint finalization and platform maturation.

Reliability

Reliability reflects the consistency of measurement across examinees. Reliability coefficients above 0.90 are generally considered strong.

Reliability can be estimated using classical methods such as Cronbach's alpha or IRT-based metrics such as Rasch real reliability. Cronbach's alpha is reported in stratified form to account for the multi-component structure of the assessment. Rasch real reliability is reported for Element 1, as it was the only portion of the exam calibrated using the Rasch model.

Table 5: Reliability Estimates (prototype vs. beta)

Metric	Prototype	Beta
Cronbach's Alpha – Full Form Stratified	0.90	0.87
Rasch Real Reliability – Element 1	0.90	0.90
Scaled Score Standard Error of Measurement (SEM) – Full Form	11.33	11.54

The beta administration demonstrated strong and stable reliability, consistent with expectations for high-stakes licensure assessment. The standard error of measurement (SEM) remained small relative to the reporting scale's standard deviation.

The SEM can be interpreted as the expected fluctuation in an examinee's observed score due solely to measurement imprecision. The observed standard error of measurement supports stable and consistent decision-making at passing score thresholds.

The replication of strong reliability under operational conditions provides support for the defensibility of bar admission decisions informed by NextGen scores.

Item and Item Set Response Time

Response-time analysis provides additional validity evidence. Extremely short response times may indicate disengagement; extremely long times may indicate confusion or poor targeting.

Median response times were computed at the item level and then averaged within each item type to reduce the influence of extreme values. The beta administration produced response-time distributions that differed from prototype expectations, with shorter response times observed for several item types.

Table 6: Average Median Response Time (minutes) (prototype vs. beta)

Item Type	Prototype	Beta
Standalone SRs	1.2	1.2
Counseling Sets	20.0	12.7
Drafting Sets	18.0	4.0
Performance Tasks	45.3	37.1

Differences in response times may reflect variation in examinee preparation or familiarity across administrations, including the longer interval between initial exposure and testing for the beta sample, as well as varied motivation for high performance on a low-stakes exam. This context is relevant when interpreting response time patterns, which do not suggest a reduction in the rigor of the assessment tasks. The absence of increased response times also suggests that blueprint finalization and UI optimization did not introduce additional cognitive burden irrelevant to the construct being measured. In some cases, reduced response times may reflect streamlined interaction with the interface or differences in how examinees engaged with the tasks.

Differential Item Functioning (DIF)

DIF analysis evaluates whether items function differently for subgroups of examinees after controlling for overall ability.

DIF analyses were conducted for gender. Sample sizes for other demographic groups were insufficient to support stable DIF detection.

Table 7: DIF Flag Rates (prototype vs. beta)

Category	Prototype % Flagged	Beta % Flagged
Selected Response: Single Select	0%	0.8%
Selected Response: Multi Select	0%	0.8%
Constructed-Response Sets	0.8%	0.8%

Flag rates during the beta administration were very low and comparable to those observed during the prototype. All flagged items underwent content review to determine whether statistical differences reflected construct-relevant factors or potential bias.

No systematic evidence of gender bias emerged.

The stability of DIF findings across administrations strengthens the fairness argument for the NextGen UBE.

Dimensionality and Local Item Dependence

Dimensionality analyses examine whether the exam measures a unified construct or multiple distinct constructs. While NextGen includes diverse item types, the intended interpretation is that all components measure minimum competence for entry-level legal practice.

Factor analytic results confirm a dominant underlying dimension consistent with the intended interpretation of minimum competence. For Element 1, the ratio of the first to second eigenvalue was approximately 4.26, and for Element 2 approximately 2.67, supporting a unidimensional structure for reporting purposes. Patterns observed for the full form were consistent with these results, with a strong first factor and a clear drop to subsequent dimensions.

Local item dependence, evaluated using Yen's Q3 statistics, remained within acceptable thresholds. Differences between maximum Q3 values and the overall average Q3 were below 0.20 for all item sets, indicating that local dependence did not meaningfully distort measurement.

Taken together, these findings support the structural validity of the reporting scale and the interpretation of a unified competence construct.

Constructed-Response Scoring

Operational Scaling and Comparative Stability

The structure of the constructed-response components of the NextGen UBE, particularly the integrated nature of the question sets, represents a central innovation of the examination model. Counseling tasks, drafting tasks, and performance tasks are designed to assess applied legal reasoning, professional judgment, and written communication in integrated contexts. Because these responses require

human scoring, the reliability and scalability of grading processes are critical to the defensibility of the examination.

The October 2024 prototype administration demonstrated that double-grading procedures could be implemented reliably at moderate scale. The January 2026 beta administration extended that work, evaluating scoring stability under larger volume and realistic timelines, using the global grading platform that will be leveraged operationally.

Beta Exam Grading Overview

The beta administration produced more than 64,000 constructed responses requiring scoring. A total of 140 graders were organized into 54 grading teams. Fourteen of the teams were supervised by a designated team leader responsible for reconciliation review and quality monitoring. Thirty of the teams worked together in consensus groups to complete the reconciliation.

Scoring followed a full double-grading model:

1. Each response was independently scored by two trained graders.
2. Scores were compared against predefined tolerance thresholds.
3. Responses exceeding tolerance limits were routed for reconciliation.
4. Team leaders conducted periodic validity seeding checks to monitor scoring consistency.

This structure mirrors the prototype model but was implemented on the digital grading platform that will be used operationally. Scaling from prototype to beta required refinement in workflow sequencing, digital scoring interfaces, and grader onboarding processes. Importantly, no structural changes were made to the scoring model itself; rather, refinements improved efficiency while preserving reliability safeguards.

The expansion from prototype volume to beta volume represents a meaningful stress test of operational readiness. The beta results indicate that scoring workflows remained stable under these expanded conditions.

Beta results also demonstrated the value of independent double grading and the effectiveness of the associated quality control procedures. Consistent with the *Standards for Educational and Psychological Testing*, scoring quality was evaluated through multiple indicators, including inter-rater agreement and the consistency of scoring decisions at key thresholds. In one instance, review of reconciliation data revealed that a grader had applied scoring criteria inconsistent with the rubric, resulting in inter-rater agreement rates between 47% and 56% across four counseling set questions. The reconciliation process identified this pattern and ensured that inconsistent scoring was corrected, thereby preserving scoring integrity and protecting examinees.

Reconciliation Rates

Reconciliation occurs when two independent graders assign scores outside an established tolerance band. Reconciliation rates provide insight into rubric clarity, grader calibration, and task stability.

Table 8: Reconciliation Rates (prototype vs. beta)

Reconciliation Rate Comparison

Item Name	Prototype			Beta			Change
	Reconciliation %			Reconciliation %			
	# Adj.	# Responses	%	# Adj.	# Responses	%	
Performance Tasks	113	2,069	5.5%	683	3,014	22.7%	17.2%
Legal Research Performance Task Short Answer	726	4,112	17.7%	187	1,501	12.5%	-5.2%
Legal Research Performance Task with One Issue	245	2,075	11.8%	126	1,501	8.4%	-3.4%
Counseling Sets	6,018	24,548	24.5%	5,107	23,155	22.1%	-2.4%
Drafting Sets	280	4,124	6.8%	104	3,240	3.2%	-3.6%

Reconciliation rates during the beta administration were comparable to or modestly lower than those observed during the prototype exam with the exception of the performance task. Slight reductions are consistent with improvements in rubric clarity and grader calibration protocols implemented after prototype feedback. The biggest change between prototype and beta grading was the implementation of teams (consensus or team-leader models). Teams that met multiple times to discuss and align on grading showed much lower rates of reconciliation than those that did not meet at all or only met at the end of grading.

From a psychometric perspective, stable or declining adjudication rates suggest that scoring variability attributable to grader disagreement did not increase under operational scaling. This finding is consistent with the scalability of the constructed-response component without evidence of degradation of scoring precision.

Inter-Rater Agreement

Inter-rater agreement (IRA) measures the percentage of responses receiving identical or tolerance-consistent scores prior to adjudication. IRA is a direct indicator of scoring reliability.

Table 9: Inter-Rater Agreement (prototype vs. beta)

Inter-Rater Agreement Comparison

Item Name	Prototype			Beta		
	Inter-Rater Agreement %			Inter-Rater Agreement %		
	#	Overall	%	#	Overall	%
Performance Tasks	1,956	2,069	95%	2,331	3,014	77%
Legal Research Performance Task Short Answer	3,386	4,112	82%	1,314	1,501	88%
Legal Research Performance Task with One Issue	1,830	2,075	88%	1,375	1,501	92%
Counseling Sets	18,530	24,548	75%	18,048	23,155	78%
Drafting Sets	3,844	4,124	93%	3,136	3,240	97%

IRA values during the beta administration remained strong and consistent with prototype findings. In several categories, modest improvements were observed, reflecting refinements in rubric specificity and calibration training.

High inter-rater agreement means that independent graders consistently reach similar conclusions about examinee performance. This consistency reduces the likelihood that scores are influenced by individual grader differences. During beta grading, teams that engaged in regular alignment discussions and systematically reviewed out-of-tolerance responses tended to demonstrate higher inter-rater agreement.

Data Analysis and Score Distributions

Constructed-response score distributions were analyzed to confirm that task difficulty remained consistent with prototype expectations. Mean raw scores, standard deviations, and distribution shapes were examined across forms.

Taken together, these results demonstrate that constructed-response scoring can be implemented reliably at scale while maintaining the controls required for high-stakes decision-making.

Grader Survey Summary

Following scoring, graders completed surveys assessing their experiences:

Clarity of Scoring Guides

Rubrics were generally viewed as clear, structured, and helpful, with an average helpfulness rating of 4.13/5.

- "The rubric was very straightforward and easy to follow."
- "Generally they were thorough and I had them in front of me 100% of the time."
- "The benchmarks and annotations oftentimes fill in the gaps when the rubric does not account for every possible permutation."

Calibration Adequacy

Seventy-three percent of the graders felt the calibration process was sufficient and that it helped maintain alignment and reduced drift.

- "This is vital to maintain the integrity of the entire process... Drift can be an issue and this helps."
- "Knowing there was ongoing calibration helped give me confidence I was staying on track (or not...)."
- "The periodic validity responses...gave me confidence that I was being consistent."

Scoring Interface Usability

Sixty percent of graders rated the platform easy or extremely easy to use, while 16% were neutral.

- "The grading platform was easy to use and I have no complaints about it."
- "The grading platform was user friendly."
- "I liked how easy it was to actually grade — read and click the radio button."

Twenty-four percent rated the platform difficult to use; the most common concerns were no ability to go back to previous responses, no ability to change the font size, and limitations in the ability to monitor quotas.

Compared to the prototype administration, beta graders reported improved clarity of rubric descriptors for the drafting and counseling sets and increased confidence in calibration materials. These qualitative findings align with quantitative reductions in adjudication rates for those item sets.

Importantly, graders indicated that the digital scoring platform was vastly improved from prototype, leading to more efficient review and less administrative burden. Platform usability is relevant to scoring reliability; complex or unstable interfaces can increase cognitive load and inadvertently introduce inconsistency.

Comparative Interpretation

Taken together, the constructed-response grading findings demonstrate

- stable or improved reconciliation rates;
- strong inter-rater agreement;
- positive grader and team leader feedback; and
- successful scaling from prototype to beta volume.

These findings confirm that the constructed-response components of the NextGen UBE maintain reliability under operationally realistic conditions. Because constructed-response scoring is often perceived as inherently more variable than selected-response scoring, this replication is particularly significant for licensure defensibility.

The beta administration therefore provides strong evidence that the integrated NextGen scoring ecosystem—content, scoring guides, training, workflow, and technology—operates cohesively and reliably at scale.

Examinee Performance

Distributional Stability and Subgroup Analysis

The evaluation of examinee performance serves two primary purposes in the context of the beta administration:

1. To confirm that overall score distributions remain consistent with prototype findings and scale expectations
2. To examine subgroup performance patterns and ensure that no new unintended disparities emerge under operationally realistic conditions

Because the prototype exam established the reporting scale and the minimally qualified candidate threshold, the beta administration provides an opportunity to evaluate whether score distributions and subgroup relationships replicate under stabilized blueprint and platform conditions.

Overall Score Analysis

Distributional Characteristics

The NextGen reporting scale ranges from 500 to 750, with a target mean of 625 and a standard deviation of 35. These parameters were established during the prototype administration and carried forward through scaling and equating in the beta administration.

The beta administration produced the following overall score distribution characteristics:

Table 10: Overall Score Summary (prototype vs. beta)

Metric	Prototype	Beta
Mean Scaled Score	625.0	617.5
Standard Deviation	35.0	32.6
Minimum Score	526	512
Maximum Score	739	737

The beta mean and standard deviation were generally aligned with prototype targets. The observed decrease in mean score and slightly lower variability are modest and within a range that does not suggest structural changes to the scale are needed.

From a measurement perspective, this stability is important. When blueprint finalization, interface changes, or item-pool expansion occur, unintended shifts in score distributions can emerge. The beta findings do not indicate meaningful distortion of the reporting scale.

Distribution Shape

Score distributions from the beta administration were unimodal, with appropriate spread across the scale. Skewness and kurtosis values were near zero, indicating an approximately symmetric distribution without excessive peakedness or tail weight. No evidence of ceiling or floor effects was observed.

This pattern is consistent with prototype results and suggests that the test continues to differentiate across the full range of ability levels relevant to entry-level legal practice.

Subgroup Score Analysis

Subgroup analysis is an essential component of assessment validation. Even when overall reliability and dimensionality are strong, differences in performance may emerge across demographic groups.

Subgroup analyses during the beta administration were conducted across key demographic categories consistent with those examined during the prototype exam.

Subgroup Differences in Item Difficulty

Table 11: Mean Item P-Values by Subgroup (beta)

Subgroup	Mean	SD	Min.	Max.
Gender				
Women	0.59	0.16	0.17	0.92
Men	0.64	0.14	0.28	0.91
Race/Ethnicity				
White	0.65	0.15	0.25	0.94
Black or African American	0.54	0.16	0.15	0.88
Asian	0.58	0.16	0.19	0.90
Latina or Latino or Latine or Hispanic	0.56	0.16	0.16	0.90
Multiracial	0.64	0.16	0.22	0.96

Within the beta administration, mean item difficulty was generally similar across subgroups, with some variation observed. For example, mean p-values were somewhat higher for men than for women and varied across racial and ethnic groups.

Observed subgroup differences in the beta administration were consistent in direction with those identified during the prototype administration. Importantly, no substantively different patterns of subgroup performance were evident in the beta sample.

In light of the DIF analyses by gender presented earlier, observed gender differences are not attributable to systematic item bias. Differences observed for other subgroups should be interpreted in the context of current sample size limitations. As larger samples become available in operational administrations, DIF analyses will be conducted for these groups to further evaluate subgroup fairness.

The replication of subgroup performance differences provides additional support for the fairness argument of the NextGen UBE and suggests that blueprint finalization and platform refinements did not introduce unintended measurement distortion.

Examinee Motivation and Readiness

One limitation of low-stakes administrations such as the prototype and beta is examinee motivation. When participants understand that their scores will not affect bar admission, performance patterns may reflect differences in engagement compared to operational testing conditions.

The January 2026 beta administration was conducted closer to anticipated launch and included participants with greater familiarity with the exam's future role. At the same time, the administration was further removed from examinees' most recent high-stakes bar examination than the prototype administration, which may have influenced preparation and readiness. Differences observed in

performance and response-time patterns across administrations are therefore best interpreted as differences in examinee engagement, rather than differences in the assessment itself.

From a psychometric perspective, variation in motivation can influence response effort and timing behavior without altering the underlying measurement properties of the exam. Response-time analyses from the beta administration indicated shorter response times across several item types, which may reflect differences in preparation, familiarity, or engagement, rather than differences in item difficulty, scoring, or construct representation.

Importantly, these patterns do not indicate structural issues with the assessment. Core indicators of measurement quality—including score distributions, item difficulty, discrimination, reliability, and model fit—remained stable across administrations. As a result, variation in examinee motivation provides context for interpreting performance and timing data but does not affect the validity of the inferences drawn from the results.

Comparative Interpretation

The examinee performance findings from the beta administration support several conclusions:

1. Overall score distributions remain stable relative to prototype scale expectations, with no evidence of distortion attributable to differences in engagement.
2. No structural shifts in the reporting scale are evident following blueprint finalization.
3. Subgroup performance patterns are consistent with prototype findings and supported by DIF screening.
4. Observed differences in examinee readiness and engagement did not produce meaningful changes in score behavior or interpretation.

Taken together, these results indicate that variation in examinee engagement did not compromise measurement stability. The reporting scale established during the prototype administration remains valid under operationally representative conditions.

The beta findings therefore provide empirical continuity across the testing arc and strengthen confidence in score interpretation for bar admission decisions.

Validation of the Recommended Passing Score Range

The recommended NextGen UBE passing score range was established through triangulation of multiple evidence streams: prototype-based scale development, statistical concordance with the legacy UBE, a national multi-method standard-setting study, and outcome-modeling analyses.

The January 2026 beta administration provided the first opportunity to evaluate that framework under operationally representative conditions. Unlike the prototype exam, the beta administration incorporated complete end-to-end processing, live platform delivery, and scaled grading workflows.

Relative to prototype results, the beta concordance study indicated a downward shift in mapped NextGen UBE scores across the recommended passing score range, with legacy-equivalent score points corresponding to lower NextGen scores. This shift reflects differences in overall performance distributions between the prototype and beta administrations and is consistent with the observed decrease in beta pass rates.

Importantly, this shift does not reflect a change in the underlying measurement model, scale, or construct being assessed. Rather, it reflects differences in examinee performance under varying levels of engagement in low-stakes conditions.

The purpose of the beta analyses was to use the data collected to evaluate three foundational elements:

1. The NextGen reporting scale
2. The recommended passing score mapping to legacy UBE scores
3. The performance of the recommended passing score range under operationally representative conditions

Validation of the Base Reporting Scale

The NextGen UBE reporting scale ranges from 500 to 750 and was initially calibrated using a two-stage hybrid scaling method anchored to prototype performance. The scale was targeted to a mean of 625 and a standard deviation of 35 to preserve interpretive continuity across administrations.

Beta data allowed evaluation of scale stability, distributional behavior, and precision at decision thresholds.

Distributional Characteristics

The beta score distribution was evaluated against projected scale parameters.

Table 12: Distributional Characteristics of Beta Scores

Statistic	Target (Prototype)	Beta Observed	Difference
Mean	625.0	617.5	-7.5
Standard Deviation	35.0	32.6	-2.4
Skewness	0.24	0.12	-0.12
Kurtosis	-0.27	0.21	0.48
% at Floor (500)	0%	0%	0%
% at Ceiling (750)	0%	0%	0%

Observed distributional parameters were generally consistent with expectations. The beta mean was modestly lower and variability slightly reduced relative to prototype targets, but these differences do not indicate structural changes to the reporting scale.

Shape characteristics of the score distribution remained appropriate. Skewness and kurtosis values were close to zero in both administrations, indicating approximately symmetric distributions without excessive peakedness or tail weight. No evidence of score compression at scale endpoints was observed.

Taken together, these findings support the stability of the reporting scale under operational conditions and indicate that the scale continues to function as intended for score interpretation.

Measurement Precision at Recommended Passing Score Range

Evaluation of measurement precision in the vicinity of recommended passing scores is critical for supporting score-based decisions. While conditional indices were not produced under the hybrid scaling approach, overall reliability and standard error of measurement (SEM) provide relevant evidence of score precision.

SEM values remained small relative to the reporting scale, and reliability estimates were consistent with prototype findings. These results indicate that measurement precision is maintained across the score scale, including in the region of typical passing scores.

These findings support the conclusion that the reporting scale provides sufficient precision to support consistent and defensible score-based decisions.

Validation of Concordance Mapping

The following analyses focus on classification behavior within the recommended passing score range. The concordance study contributed to the development of the recommended mapping between legacy UBE passing scores (260–270) and NextGen UBE scores (610–620). This mapping was derived from examinees who completed both the July 2024 legacy UBE and the October 2024 NextGen prototype administration, using equipercentile linking methods that were evaluated for consistency and interpretability.

As observed in the prototype analyses, concordance provides a basis for relating performance across exams but is not intended to produce exact prediction of individual outcomes. Instead, it supports consistent interpretation of scores across testing programs and administrations.

The January 2026 beta administration provided the first opportunity to evaluate whether this mapping remains appropriate under operationally representative conditions. Compared to prototype results, the beta concordance analysis showed modest deviations from one-to-one mapping within the recommended score range, including isolated score gaps and many-to-one mappings.

These deviations are expected when applying concordance methods across administrations with differing performance distributions. Importantly, they do not materially affect interpretation within the recommended passing score range, which is supported by multiple sources of evidence beyond concordance alone.

Pass-Rate Behavior at Passing Standards in the Recommended Range

Observed beta pass rates were calculated at scaled score points within the recommended NextGen UBE passing score range (610–620) and compared to prototype-based pass rates derived from outcome analysis.

Prototype analyses indicated that pass rates were responsive to relatively small changes in score thresholds, with incremental increases in passing scores producing meaningful decreases in pass rates across this range. In addition, pass rates for the prototype sample were generally lower than those observed in operational administrations, likely reflecting differences in examinee readiness and motivation.

Table 13: Observed Pass Rates at Recommended Passing Standards (prototype vs. beta)

Legacy Reference Score	NextGen Score	Prototype Observed Pass Rate	Beta Observed Pass Rate	Difference
260	610	65.5%	61.0%	-4.5%
261	611	63.8%	59.4%	-4.4%
262	612	62.4%	58.1%	-4.3%
263	613	60.7%	56.5%	-4.2%
264	614	59.6%	55.4%	-4.2%
265	615	58.2%	54.0%	-4.2%
266	616	57.3%	52.4%	-4.9%
267	617	56.2%	51.2%	-5.0%
268	618	55.3%	50.1%	-5.2%
269	619	54.2%	49.2%	-5.0%
270	620	52.9%	47.5%	-5.4%

Prototype pass rates decreased monotonically across the recommended score range, from approximately 65.5% at 610 to 52.9% at 620, illustrating the expected sensitivity of classification outcomes to relatively small changes in the passing score.

Observed beta pass rates followed a similar monotonic pattern but were lower across the range, with differences of approximately four to five percentage points. This shift may reflect differences in examinee readiness, timing, or engagement between the prototype and beta administrations. This pattern is consistent with the downward shift observed in concordance mapping, in which equivalent legacy UBE score points correspond to lower NextGen UBE scores in the beta administration.

Consistent with the concordance analyses, minor irregularities in the mapping relationships were observed in the beta administration, including isolated score gaps and many-to-one mappings within the recommended range. Despite these localized deviations, the overall relationship between score thresholds and classification outcomes remains consistent across administrations.

Stability Under Difficulty Variation

Form-level difficulty differences between prototype and beta administrations were evaluated in relation to classification outcomes.

Results indicate that these observed differences in difficulty do not materially alter the relationship between passing-score points and classification outcomes.

Validation of the Recommended Passing Score Range

The recommended passing score range reflects convergence across standard-setting panel judgments, concordance results, and outcome modeling. Beta data allowed examination of classification behavior in the region most relevant to jurisdictional policy.

Synthesis

The beta administration provided additional evidence supporting the framework established through prototype-based research. Findings across score distributions, measurement precision, concordance results, pass-rate behavior, and subgroup impact analysis are consistent with expectations for an operationally representative administration.

Beta findings support three core conclusions:

1. The NextGen reporting scale functions as intended under operational conditions.
2. The relationship between legacy UBE and NextGen UBE passing standards remains appropriate for supporting score interpretation, although beta results indicate a systematic downward shift in mapped NextGen scores relative to prototype findings.
3. The recommended passing score range produces classification outcomes consistent with prototype-based expectations and historical patterns, despite a downward shift in mapped NextGen scores observed in the beta administration.

These findings from the beta administration reinforce that the recommended passing score range is supported by multiple sources of evidence and remains appropriate for informing jurisdictional decision-making during the transition to the NextGen UBE.

Summary of the Psychometric Evidence and Passing Score Validation

Across the testing arc, the NextGen UBE has demonstrated stable and consistent psychometric performance under increasingly operational conditions. The January 2026 beta administration replicated key findings from the October 2024 prototype exam across item difficulty, discrimination, dimensional structure, and reliability, confirming that measurement properties remain within expected ranges for high-stakes licensure assessment.

Critically, these results were obtained under fully integrated, operationally representative conditions, including finalized blueprint specifications, scaled grading processes, and live administration environments. This replication strengthens confidence that the observed measurement characteristics are not artifacts of controlled testing conditions, but are stable under real-world administration.

The recommended passing score range was established through multiple converging lines of evidence during the prototype phase, including standard-setting judgments, concordance analyses with the legacy UBE, and outcome-based validation. The beta administration provides confirmatory evidence that the underlying scale, score distributions, and measurement precision supporting that recommendation remain stable.

Reliability estimates, standard error of measurement, and item-functioning patterns observed in the beta administration indicate that score interpretations at decision thresholds are consistent and defensible. No evidence emerged of systematic shifts in difficulty, discrimination, or subgroup performance that would call into question the established scale or passing standard.

Taken together, the evidence supports the conclusion that the recommended passing score range remains valid, stable, and appropriate for operational use.

Part III Summary: Testing Arc Synthesis

Part III documents the staged empirical program used to develop and confirm the NextGen UBE as an integrated licensure assessment system. Across pilot, field test, prototype, and beta phases, NCBE evaluated item family functioning, structural coherence, scoring reliability, fairness indicators, and operational robustness under progressively more realistic delivery conditions.

The pilot phase served as the primary construct-refinement stage. Early analyses established that the item families elicited the intended integration of knowledge and skills, that draft scoring materials differentiated performance levels, and that timing and interface assumptions could be tuned before large-scale administration.

The field test expanded scale and heterogeneity of the participant pool and generated early technical evidence on measurement targeting, response time behavior, and analytic grading. Difficulty distributions were appropriately centered for licensure decision-making. Timing analyses did not

indicate systemic speededness, and grading results supported the viability of an absolute analytic scoring model with double grading. Early subgroup analyses produced no evidence that the new item formats materially amplified performance gaps beyond patterns typical in licensure testing.

The prototype administration represented the first full-form psychometric evaluation. It established the reporting scale, calibrated items using Rasch-family models, and confirmed reliability and structural coherence under full-length conditions. Dimensionality analyses supported a single composite score interpretation, with expected local dependence patterns attributable to item set structure. DIF screening did not reveal systematic subgroup bias; DIF analyses will continue to be conducted as larger samples become available in operational administrations. The prototype also produced foundational policy-facing outputs, including national standard-setting evidence, relationship to the legacy UBE (informed by the concordance study), and outcome modeling supporting a recommended passing score range.

The beta administration functioned as the operational confirmation phase. Unlike earlier administrations, beta was designed to evaluate the complete ecosystem end to end—including readiness workflows, live delivery stability, response preservation and recovery, scaled grading operations, and multi-site jurisdictional monitoring—under launch-ready conditions. From a measurement standpoint, the beta provided the first operationally representative dataset with which to replicate prototype psychometric findings, confirm stability of item and score distributions across forms, and evaluate the behavior of the reporting scale and decision region under realistic examinee engagement and administration constraints.

Finally, beta analyses were used to validate—not re-derive—the recommended passing score range framework by stress-testing: (1) reporting scale stability and precision in the 610–620 region, (2) concordance consistency relative to legacy-equivalent cut-score points, and (3) pass-rate and subgroup sensitivity across the recommended range. The tables in the preceding section identify the specific metrics used for this validation.

Part IV. Operational Readiness

The development of the NextGen UBE was intentionally structured as a staged empirical program. Each phase of the testing arc—pilot testing, field test, prototype exam, and beta administration—served a distinct purpose in evaluating construct representation, measurement stability, scoring reliability, fairness indicators, and delivery integrity. Part IV synthesizes the accumulated evidence to assess whether the examination is ready for operational implementation and to define the framework that will govern post-launch monitoring.

Operational readiness in high-stakes licensure assessment requires more than acceptable reliability coefficients or positive user feedback. It requires convergence across psychometric, operational, technological, and governance domains. The examination must measure the intended construct, classify examinees consistently at decision thresholds, function without delivery distortion, and maintain fairness across demographic groups. It must also operate cohesively as a system rather than as isolated components.

The evidence documented in this report demonstrates that those conditions have been met.

Replication and Stability Across Administrations

A central principle of defensible measurement is replication.

The prototype administration established the NextGen reporting scale, calibrated items under a Rasch-family modeling framework, evaluated dimensional structure, conducted national standard setting, and performed concordance analyses with the legacy UBE. The beta administration was designed not to generate new psychometric theory but to confirm that those findings persist under operationally representative conditions.

The beta results demonstrate that the structural properties observed during prototype remain stable. Item difficulty distributions replicate within expected tolerance bands. Discrimination indices remain within acceptable ranges and show no systematic degradation following blueprint finalization. Reliability estimates continue to meet high-stakes licensure standards. Dimensionality analyses support interpretation of a unified composite score representing minimum competence for entry-level legal practice.

Importantly, the beta administration did not introduce structural distortion in score distributions. Means and standard deviations aligned with projected scale parameters. No evidence of compression at the scale floor or ceiling emerged. Overall measurement precision, including in the decision region corresponding to the recommended passing score range, remained stable. These findings confirm that the reporting scale functions as intended when applied to a fully integrated, operationally realistic administration.

Scoring Integrity at Scale

The NextGen constructed-response scoring ecosystem represents a central innovation of the NextGen UBE.

The transition to analytic rubrics, independent double grading, structured adjudication workflows, and digital scoring platforms required careful evaluation under increasing volume.

The beta administration extended prototype scoring evidence by stress-testing grading workflows at substantially greater scale. Inter-rater agreement remained strong. Adjudication rates were consistent with, and in several cases modestly lower than, those observed during the prototype administration. Score distributions for constructed-response components remained stable across forms. No evidence suggests that scaling introduced additional variability or compromised scoring precision.

Equally important, the grading platform operated with consistent routing logic, audit trails, and reconciliation workflows. These controls are integral to licensure defensibility, as they ensure that scoring decisions are traceable, reviewable, and governed by predefined tolerance thresholds. The beta administration confirms that the scoring model is scalable without loss of integrity.

Delivery Integrity and System Cohesion

High-stakes examinations must preserve response integrity under real-world conditions.

Psychometric properties are meaningful only if delivery conditions do not introduce construct-irrelevant variance.

The beta administration incorporated full-system delivery using the finalized digital ecosystem, including candidate readiness workflows, jurisdiction monitoring interfaces, secure delivery applications, failover recovery mechanisms, and integrated data flows.

Intentional failover simulations confirmed that response preservation protocols function as designed. No response data were lost during simulated interruptions. Completion rates exceeded 99 percent, and no examinee failed to complete due to technology instability. Multi-site administration across jurisdictions did not reveal form-level anomalies or site-specific distortions.

These findings are operationally significant and psychometrically consequential. Stable delivery conditions protect comparability across administrations and ensure that score variation reflects examinee ability rather than technological artifact.

Fairness and Subgroup Stability

Across the testing arc, subgroup performance patterns have remained consistent with historical licensure testing patterns.

Observed subgroup differences in mean item p-values during the beta administration were consistent in magnitude and direction with prototype findings and did not indicate the emergence of new structural disparities.

The replication of fairness indicators under operational conditions strengthens confidence that blueprint stabilization and ecosystem maturation did not introduce unintended measurement distortion.

Validation of the Passing Score Framework

The recommended passing score range was derived from a synthesis of standard-setting judgments, a statistical concordance study, and outcome modeling.

The beta administration provided the first opportunity to evaluate the behavior of that framework under operationally realistic conditions.

Beta findings confirm that the reporting scale remains stable, that overall measurement precision remains within acceptable bounds, and that pass-rate behavior across the 610–620 range is consistent in pattern with prototype results.

No evidence of discontinuities in classification outcomes emerged across mapped legacy-equivalent cut scores.

These results provide empirical support for the continued use of the recommended passing score range during early operational administrations, subject to each jurisdiction's independent authority to establish its own standard.

Risk and Ongoing Oversight

Operational readiness does not imply that the system will remain static.

It requires a defined framework for ongoing monitoring. Early operational administrations will include structured review of distributional stability, inter-rater agreement metrics, adjudication rates, DIF flag rates, and platform performance indicators. These metrics will be evaluated relative to prototype and beta benchmarks to ensure continuity.

Where deviations emerge, the governance structure established during development will support timely review and adjustment. Continuous monitoring is an integral component of responsible licensure testing and will remain embedded within the NextGen operational model.

Conclusion

The NextGen UBE represents a measured evolution of licensure assessment—one grounded in evidence, shaped by stakeholder input, and tested under conditions that reflect real-world administration.

Across multiple phases of development, NCBE has evaluated not only individual components of the exam, but the performance of the full system: content, platform, administration, and scoring. The evidence presented in this report demonstrates that these components function together as intended, producing stable, reliable, and defensible measures of minimum competence.

The psychometric results are consistent and replicable. The operational model has been tested at scale. The digital platform has demonstrated stability, accessibility, and resilience under live conditions. Scoring processes have been implemented with the controls and oversight required for high-stakes decision-making.

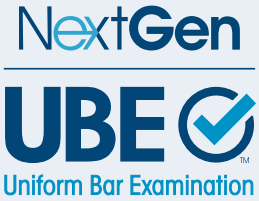
Importantly, these outcomes have been observed not in isolation, but through an integrated testing arc designed to reduce uncertainty at each stage of development.

Innovation in licensure assessment carries inherent responsibility. New item types, digital delivery, and expanded measurement of skills introduce complexity—but they also provide an opportunity to better align licensure testing with the realities of modern legal practice. Throughout this process, NCBE has approached that responsibility conservatively, ensuring that each advancement is supported by empirical evidence and does not compromise the standards that jurisdictions rely on.

The conclusion is not that the exam is simply different. The conclusion is that it works.

It works as a measurement system. It works operationally. And it works in service of the core purpose of licensure: protecting the public by ensuring that newly licensed lawyers are prepared for safe and effective practice.

The NextGen UBE is ready for operational launch.


NextGen
UBE
Uniform Bar Examination


NCBE
National Conference
of Bar Examiners

ncbex.org