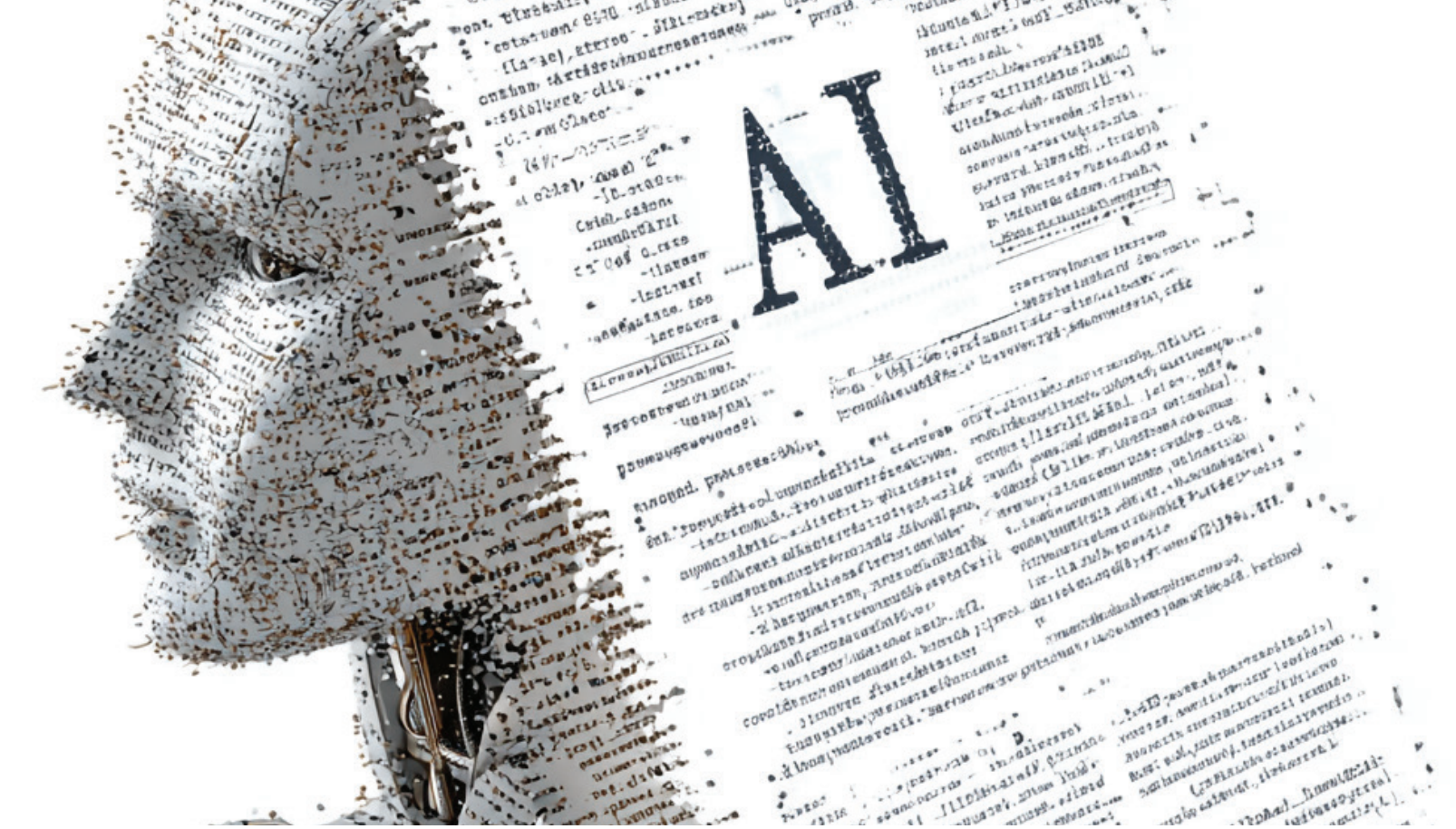

ARTIFICIAL INTELLIGENCE AND STATE ASSESSMENT CONTRACTS:

Translating Technical Guidance into Clear Language

March 2026

André A. Rupp

Center for Assessment

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

National Center for the Improvement of Educational Assessment, Inc. (the Center for Assessment) is a New Hampshire based not-for-profit (501(c)(3)) corporation. Founded in September 1998, the Center's mission is to improve student learning by partnering with educational leaders to advance effective practices and policies in support of high-quality assessment and accountability systems. The Center for Assessment does this by providing services directly to states, school districts, and partner organizations to support state and district assessment and accountability systems.

ACKNOWLEDGEMENTS:

Many thanks to Scott Marion for the initial guidance in developing this document and to Ruhan Circi for engaging in a consulting conversation and sharing some of her materials for additional inspiration. Thanks also to Will Lorié and Nathan Dadey for sharing an advance copy of their framework paper on artificial intelligence in large-scale assessment. I also want to thank the members of the Maryland Technical Advisory Committee who engaged in a critical discussion during the December 2025 meeting and provided actionable feedback.

Portions of the analysis and illustrative contract language were developed with the assistance of Perplexity Pro (Perplexity AI). The author retains full responsibility for the content of this document.

SUGGESTED CITATION:

Rupp, A. A. (2026). *Artificial intelligence and state assessment contracts: Translating technical guidance into clear language.* Dover, NH: The National Center for the Improvement of Educational Assessment.

PHOTO CREDIT:

Created with Midjourney.

TABLE OF CONTENTS

ARTIFICIAL INTELLIGENCE AND STATE ASSESSMENT CONTRACTS:

Translating Technical Guidance into Clear Language

OVERVIEW

This brief lays out key opportunities and risks for utilizing artificial intelligence (AI), in particular generative AI (GenAI), for different phases of a statewide assessment program.

It then translates these into considerations about the types of contract language that should be included in requests for proposals (RFPs) and vendor contracts to ensure that the vendor takes proper steps to safeguard against the risks. As suggested in this brief, these could be bundled into a separate AI addendum in the master contract.

The brief is organized into three major components:

- General frameworks that can provide guidance on understanding the technical quality and necessary safeguards for AI/GenAI systems (Section 1)
- Cross-cutting governance principles are required to ensure transparency, oversight, security, and proper system maintenance for AI/GenAI systems (Section 2)
- Full phase-by-phase assessment lifecycle analysis with potential applications, risks, and contract levers (Section 3)

Sections 2 and 3 can be seen as complementary entry points for working through key AI/GenAI considerations, with Section 2 walking through these considerations by type and Section 3 walking through them by assessment phase.

The appendix presents potential contract clauses that operationalize these principles. As we are not lawyers, this language should be viewed as general guidance rather than official legal text.

Depending on perceived risk, not every clause from this brief will need to be included in the final contract, nor is this an exhaustive list. At a minimum, a state department should ensure that each of the cross-cutting governance themes in Section 2 is represented in some form.

It is important that experts in the assessment teams in the department work closely with the procurement office in this process for the mutual benefit of all involved parties.

SECTION 1 - GENERAL FRAMEWORKS

In this section we discuss terminology, technical quality frameworks, and professional standards in the field to provide a general context for the subsequent discussions.

1.1 Terminology

For purposes of this brief and any associated RFP or contract language:

- Vendor is any organization contracted by the state department to provide assessment services, including development, administration, scoring, reporting, or related AI/GenAI-enabled services.
- AI system is any engineered or machine-based system that uses data to generate outputs such as predictions, classifications, recommendations, or content that influence any part of the assessment lifecycle.
- Generative AI refers to AI systems that are designed to generate substantially new, human-interpretable content (e.g., free-form text, code, images, or explanatory narratives) in response to prompts or inputs, using probabilistic, machine-learned, or large language models.
- AI/GenAI component is a specific model, service, or module that implements AI or GenAI functionality within a larger process (e.g., item generation, scoring, anomaly detection, or report drafting).
- AI/GenAI-assisted process is a process in which AI/GenAI components support human work but do not themselves have final authority over high-stakes decisions (e.g., AI/GenAI-generated item drafts that are always reviewed by human committees).
- AI/GenAI-related incident refers to any event in which an AI component behaves in a way that materially deviates from its documented design or expected performance, including but not limited to biased outputs, misleading or hallucinated content, security anomalies, or unanticipated effects on scores, classifications, or reporting.

More specifically, routines that simply automate tasks are generally not considered AI, and certainly not GenAI. However, given that automated scoring and routing have such a prominent place in large-scale assessment it is included in this brief under AI/GenAI for simplicity. Similarly, this document uses the term “statistical” to refer to social-science-based and psychometric approaches.

ATP has released a free glossary with terms for AI/GenAI in assessment that provides more differentiated definitions, while Johnson (2025) includes a similar glossary at the end of the research report. Various other glossaries are publicly available.

1.2 Technical Quality Frameworks

Table 1 presents an illustrative selection of relevant AI/GenAI frameworks for large-scale assessment in K-12. They differ in relative emphasis but generally include areas of mutual overlap to support responsible uses of AI/GenAI. New frameworks and landscape analyses are released frequently.

These frameworks provide guidance on the considerations that specialists working with AI/GenAI technologies need to bear in mind in their work. Thus, they can be viewed as reflecting expectations that the state department should have for the vendor it selects for assessment.

Various other national and international frameworks that articulate high-level principles exist and are too numerous to list here. However, they typically intersect only in specific subsections of assessment-specific frameworks (e.g., governance or transparency). Examples include the OECD AI Principles, the EU AI Act, and the US AI Bill of Rights.

1.3 Professional Standards for Educational Assessment

Professional standards for educational assessment are their own kind of framework, although they differ in rhetorical orientation and practical use from some of the other general frameworks in Table 1. They generally have an authoritative tone and describe minimum expectations for certain areas of educational assessment work.

Moreover, audit processes for educational assessments with their associated rubrics and protocols, including federal peer review, are generally aligned with such standards, which serve as their conceptual spine.

Specifically, the Standards for Educational and Psychological Testing are currently being revised and will include various guardrails for the responsible use of AI/GenAI. Similarly, the ITC and ATP Guidelines for technology-based assessment discuss various uses of AI/GenAI, in particular in section IV. Buros maintains a repository of core standards documents relevant to educational and psychological testing.

Whenever new authoritative standards are officially published, vendors should review their internal practices for alignment and make adjustments as needed. State departments should similarly review them to ensure that their current practices and expectations for the vendor are up to date. If there are critical gaps, it is crucial to discuss them with the vendor in a timely manner and find resolutions to close them.

Table 1. Select Frameworks for AI/GenAI in Educational Assessment

Reference	Description
Association of Test Publishers (May 2025). Considerations for the Use of Generative AI in Test Development: A Special Publication from ATP. ATP. [behind firewall]	An integrative framework that walks through considerations of AI/GenAI technologies for different aspects of test development writ large, and a strong emphasis on risk, quality, ethics, and test integrity.
Bulut, O., et al. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. Chinese/English Journal of Educational Measurement and Evaluation, 5(3), https://doi.org/10.59863/MIQL7785	A discussion paper around ethical implications of AI-powered tools in educational measurement around issues such as automation bias and environmental impact with proposed solutions to ensure AI's responsible and effective use in education.
Burstein, J., LaFlair, G. T., Yancey, K., von Davier, A. A., & Dotan, R. (2024). Responsible AI for test equity and quality: The Duolingo English Test as a case study. arXiv.	A high-level framework for responsible AI use is illustrated with the Duolingo assessment program, with sample descriptions/explanations for key claims.
European Union (2023). Proposal for standard contractual clauses for the procurement of Artificial Intelligence (AI) by public organisations. European Union.	A collection of example clauses for general contracts around AI technologies, which exist separately for low-risk and high-risk applications.

Reference	Description
Johnson, M. S. (2025). Responsible AI for measurement and learning: Principles and practices (Research Report No. RR-25-03). Princeton, NJ: Educational Testing Service.	Integrative framework grounded in international research encompassing fairness and bias mitigation, privacy and security, transparency, explainability, accountability, educational impact, and integrity, and continuous improvement.
Lorié, W., & Dadey, N. (2026). Artificial intelligence in large-scale assessment programs: Applications and considerations for state education agencies. Dover, NH: Center for Assessment.	An integrative framework that describes the opportunities, risks, and current state of knowledge for the use of AI/GenAI across the different phases of large-scale assessment development.
National Institute for Standards and Technology (2023). Artificial intelligence risk management framework (AIRMF 1.0). NIST.	Integrative framework describing ways to think about and operationalize risk around AI/GenAI systems, which also includes a comprehensive practical "playbook" with additional details, examples, and descriptions.
Rafal-Baer, J. and Smith, L. (2024, June). Framework for implementing artificial intelligence (AI) in state education agencies (SEAs) (Version 1.0). ILO Group.	An integrative framework describing the various uses of AI/GenAI for state education agencies, which includes a variety of cross-links to other frameworks for the responsible use of AI/GenAI.

In the following section we walk through different cross-cutting considerations for the design, implementation, and monitoring of AI/GenAI tools across the assessment lifecycle (e.g., transparency, fairness, risk assessment). In the section after that, we present these considerations separately for each assessment lifecycle phase (e.g., item development, scoring, reporting).

Thus, both sections should be viewed as complementary entry points for determining the key considerations that need to be addressed for a given assessment program and how they can be best represented through contract language.

SECTION 2 CROSS-CUTTING GOVERNANCE THEMES

As illustrated by the frameworks cited in the previous section, professional quality considerations around AI/GenAI are essentially designed to ensure that information from these systems meets essential standards for core technical quality criteria (e.g., accuracy, validity, fairness, transparency, security).

Consequently, many of the considerations for the use of AI/GenAI in different phases of the assessment lifecycle are very similar across different frameworks at a higher level of grain size. The following considerations can thus be found in multiple frameworks with different frameworks often emphasizing, and going deeper into, specific aspects.

In the following, we briefly describe the motivation around core considerations and then provide guidance around the kind of information that vendors should generally disclose. These should be

seen as representative of core types of information but should be adapted to the specific needs and use contexts of a given state education agency and assessment program.

2.1 Transparency & Explainability

General transparency and explainability, within reasonable guardrails, is foundational for ensuring an effective understanding of how systems are designed, perform, and could be improved.

While vendors can be expected to maintain certain computational components behind their internal firewalls (e.g., training data, code suites, server access), they should be expected to share many other aspects with clients.

In many ways, this is about educating state education agency staff so that their internal teams of specialists are empowered to ask critical questions about the technology. This, in turn, ensures that agency staff can make informed internal decisions about the pros and cons of the particular assessment solution(s) offered by a vendor.

Vendors should disclose the following technical information:

- A complete inventory of all AI/GenAI components / tools used in assessment, including each component's purpose, inputs, outputs, and where it sits in the assessment workflow (e.g., content development, routing, scoring, translation, accessibility supports, data forensics, and reporting).
- The specific ways each AI/GenAI component affects relevant assessment aspects, with a clear differentiation of risk levels. These aspects include:
  - student score generation (e.g., automated scoring, task generation).
  - score interpretations (e.g., customized on-demand reports, explanations, or exemplar items).
  - process efficiencies (e.g., scheduling optimizations, remote proctoring).
  - technical support (e.g., administration guidance, resolution of administration issues)
  - validation processes (e.g., creation of technical report segments, design of research plans or protocols)
- High-level computational specifications used to develop and evaluate the AI/GenAI algorithms (e.g., initial training data, fine-tuning data, model families, performance metrics, evaluation protocols, and hosting environments).
- High-level descriptions of known limitations of each AI/GenAI component (e.g., hallucination risks, error modes).
- The roles and contributions of subject-matter experts, psychometricians, and other stakeholders in designing, validating, and monitoring AI/GenAI systems.
- Documentation describing how AI/GenAI uses comply with applicable federal and Maryland laws, regulations, and relevant professional standards, and how compliance is monitored over time.

Vendors should also provide the following rollout supports:

- Clear descriptions of risks to validity, fairness, security, and data quality introduced by AI/GenAI, and the controls used to mitigate those risks.
- Plain-language explanations, FAQs, and technical summaries that the state department can share with educators, families, and the public to explain where AI is used and how it is overseen.

- Explicit labeling of AI/GenAI-generated or AI/GenAI-assisted content in materials provided to the state department and, where appropriate, in professional-learning resources.
- Training and reference materials for the state department staff that describe appropriate and inappropriate uses of AI/GenAI outputs, escalation paths, and override procedures.
- Support to the state department in preparing clear, accessible communication materials in the event of AI/GenAI-related incidents that materially affect student scores, classifications, reporting, or test security, including draft talking points and FAQs that the state department can adapt for different audiences for such situations.

2.2 Bias & Fairness Investigations

The issues of bias and fairness are technically aspects of performance evaluations mentioned under 2.1, but they are called out here separately due to their particular importance in educational assessment.

Without careful training, evaluation, and ongoing oversight, AI/GenAI systems are likely to create outputs that are biased in ways detrimental to the assessment enterprise, which may lead to unfair interpretation and uses of assessment information (e.g., discriminatory practices). This may include biases for students from different demographic groups, service-based groups, or groups with unique sociocultural or linguistic backgrounds.

Bias and fairness investigations are particularly important for AI/GenAI systems that are updated frequently, with each new version introducing powerful capabilities but also risks such as previously unseen types of bias. In practice, fairness and accuracy often exist in tension, and an acceptable compromise needs to be found given the stakes associated with score use.

Vendors should provide:

- A documented framework for identifying and evaluating bias and fairness in AI/GenAI systems, including definitions of key terms, specifications of student groups, and relevant assessment aspects for which these issues are particularly salient (e.g., adaptive presentations and scoring of content, accessibility supports).
- A description of statistical models, metrics, benchmarks, and qualitative review processes used to assess bias at development and at each major model update.
- Evidence that AI/GenAI components deployed in operational use meet predefined quality-control thresholds and that any residual disparities continue to be monitored, documented, and addressed through mitigation plans agreed upon with the state department.

2.3 Human-in-the-Loop Oversight

AI/GenAI does not replace psychometric, content, or policy expertise. Human-in-the-loop (HITL) oversight ensures human authority in all critical areas, thus strongly reducing risks that could arise from too much human-AI dependence without critical engagement.

Vendors should ensure that:

- Human committees or designated state department authorities retain final approval over all assessment content, including items drafted or revised using AI/GenAI.
- AI/GenAI-generated or AI/GenAI-assisted scoring recommendations, security flags, or report content do not take effect without human review and explicit acceptance in defined decision contexts.

- HITL checkpoints, escalation criteria, and override mechanisms are documented, with associated rationales, evidence logs, and audit trails available to the state department.
- The state department has the right to audit AI/GenAI-assisted processes and outputs and to require adjustments to HITL configurations that impact critical quality aspects such as validity, fairness, or policy alignment.

2.4 Security & Privacy

AI/GenAI introduces multiple new areas with security risks that need various safeguards to ensure reliable use, many of which are extensions of safeguards that existed before AI/GenAI was explicitly used. Ideally, the development of AI/GenAI systems follows a security-by-design approach. This is similar to the idea of accessibility-by-design where the issue is considered natively throughout the development, implementation, and evaluation process.

Vendors should ensure that:

- The state department retains ownership of all student-, school-, and state-level assessment data, and the vendor's rights to use those data are limited to fulfilling contract obligations and explicitly approved model improvement activities.
- All AI/GenAI data flows and processing activities comply with applicable federal and state privacy and security laws, as well as the state department policies, including limitations on using student data for general model training.
- Student responses, secure items, and related assessment artifacts are never transmitted to unapproved public AI services, external APIs, or other uncontrolled third-party endpoints.
- Models are hosted on secure infrastructures that implement strong access controls, encryption, monitoring, and logging to prevent unauthorized access or data leakage.
- AI/GenAI-specific penetration tests, adversarial testing, and red-team exercises are conducted regularly, with remediation plans and timelines for any identified vulnerabilities.
- Security configurations and controls are reviewed and updated in a timely manner to address newly emerging threats without disrupting operational assessment activities.

2.5 System Maintenance & Governance

AI/GenAI models are updated frequently, which can affect performance across different assessment development phases and the associated interpretational validity or scores. While ongoing vendor innovation is needed to improve systems, the rollout of these innovations into operational practice requires care. Moreover, AI/GenAI systems are computationally complex and require effective internal data architecture, engineering, and governance practices.

Vendors should ensure that:

- The state department has prior review and approval authority for foundational changes to AI/GenAI components that materially affect scoring, routing, reporting, security, or other critical assessment functions.
- Version histories, update rationales, validation evidence, and change-impact analyses are maintained for all AI/GenAI components and shared with the state department upon request.
- Data governance policies and practices are documented and implemented consistently, including data quality controls, lineage tracking, retention schedules, and access management.

- The vendor maintains sufficient internal expertise, staffing, and infrastructure to support responsible development, monitoring, and continuous improvement of AI/GenAI systems over the life of the contract.

2.6 Testing, Evaluation, and Monitoring Expectations

For each phase of the assessment lifecycle in which AI/GenAI systems or AI/GenAI-assisted processes are used the vendor should describe:

- Pre-deployment readiness checks, including test sets, metrics, and acceptance thresholds appropriate to each lifecycle phase.
- Quantitative evidence that will be provided to the state department during operational use (e.g., customized session reports for scoring batches) and after the testing windows have closed (e.g., for inclusion in technical reports and peer-review documentation).

These expectations apply, at a minimum, to AI/GenAI-assisted construct and blueprint development, item and task development, field testing and equating, administration and test security, response scoring, and reporting and interpretation.

Special attention should be paid to operational monitoring procedures for automated scoring processes, which requires documentation of the frequency of reviews, types of subgroup analyses where applicable, quality of human scoring, and threshold-based triggers for escalation or mitigation.

2.7 AI Risk Management Requirements

Risk management is an effective cross-cutting component of the work with AI/GenAI systems and complex technological systems more generally. To support effective use of AI/GenAI across assessment, the vendor shall assist the state department in implementing a comprehensive risk-management approach that spans governance, lifecycle evaluation, and continuous improvement.

Specifically, the vendor should:

Support the state department's AI risk governance activities

- Provide documentation, data, and expert consultation needed for the state department to define, monitor, and periodically update AI/GenAI-related risk expectations for assessment, including distinctions between high-stakes and low-stakes uses and between pilots and fully operational deployments.
- Participate in the state-department-led AI/GenAI risk reviews at least annually, including discussion of risk tolerances, emerging risks, and planned mitigations for AI/GenAI components.
- Address safety, trustworthiness, and residual risk explicitly in reports for pre-operational readiness checks or operational monitoring evaluations
- Identify potential safety-related harms from AI/GenAI components (e.g., misleading score interpretations, inappropriate security flags, or biased content) and describe safeguards, fail-safe behaviors, and conditions under which AI/GenAI-enabled functions will be disabled, constrained, or rolled back.
- Document known limitations and residual risks for each material AI/GenAI component (including scoring, content generation, monitoring, and reporting) and provide the state department with updated residual-risk summaries after each major model update.

Provide an AI/GenAI-specific test and evaluation plan

- Develop and maintain a written plan describing how AI/GenAI inputs, models, and outputs will be tested, evaluated, and monitored across the assessment lifecycle, including planned metrics, test sets, subgroup analyses, and acceptance thresholds.
- Share concise test and evaluation summaries with the state department in a form that can be incorporated into technical documentation and peer-review evidence, including descriptions of uncertainty, limits of generalization, and monitoring triggers.

Ensure effective human–AI interaction and operator competency

- Provide training, reference materials, and decision-support guidance for the state department and vendor staff who use AI/GenAI-assisted tools (e.g., scoring dashboards, anomaly flags, or AI/GenAI-generated reports), emphasizing appropriate reliance, escalation paths, and override procedures.
- Evaluate usability of AI/GenAI-assisted tools used by educators, scorers, or the state department staff and incorporate feedback from the state department to reduce misinterpretation or over-reliance on AI/GenAI outputs.

Facilitate stakeholder engagement and impact assessment

- Support the state department in gathering and summarizing feedback from districts, educators, families, students, and relevant stakeholder groups regarding AI/GenAI use in assessment, and propose adjustments to AI/GenAI configurations or processes where needed.
- Conduct and provide periodic AI/GenAI impact assessments that synthesize evidence on validity, fairness, privacy, security, and educational impact by relevant student subgroups, including recommended and implemented mitigations.

Implement AI incident management and decommissioning procedures

- Maintain and follow an AI/GenAI-specific incident response process covering issues such as scoring anomalies, hallucinated or misleading reporting content, problematic security flags, or regressions introduced by model updates, including timelines for detection, investigation, remediation, and communication with the state department.
- Define and document decommissioning and rollback procedures for AI components (including handling of training and finetuning data and protection of longitudinal score comparability) and execute such procedures when the state department determines that an AI component is no longer acceptable for use.

Manage third-party and supply-chain AI risks

- Identify all third-party AI/GenAI services, models, or libraries used in assessment processes and provide documentation of their roles, dependencies, and available assurance or risk-management evidence.
- Establish contingency plans for critical third-party AI/GenAI components whose behavior, availability, or terms may change, to ensure assessment validity, fairness, and security are maintained without interruption.

2.8 State Department Rights

The state department should retain final decision-making authority over whether and how AI/GenAI systems are used in assessment, which includes:

- Determining whether AI/GenAI is used at all for a given function such as blueprint development, item design, scoring, test security, or reporting.
- Determining the conditions under which AI components may be piloted, limited to low-stakes contexts, or used in fully operational, high-stakes settings.
- Retaining final approval for operational use of foundational changes to AI components that materially affect any of the key phases of the assessment lifecycle, which includes specifying timelines for implementation and any required transition or parallel-run periods.

Table 2 on the next page summarizes the key ideas from this section in a compact form. In the next section of this brief, we revisit the ideas in this section by going through all of the core phases of the assessment lifecycle.

Table 2. Summary of Cross-cutting Governance Themes

Aspect	Focus	Vendor Expectations
Transparency & Explainability	Developing AI/GenAI components, their roles, limitations, and compliance posture understandable to technically informed state staff and external audiences.	Full inventory of AI/GenAI components, descriptions of data, models, and limitations; plain-language explanations; labeling of AI-generated content; training and FAQs for staff.
Bias & Fairness	Ensuring AI/GenAI does not introduce or amplify inequities for student groups across assessment uses.	Documented fairness framework; metrics and subgroup analyses; evidence of meeting thresholds; ongoing monitoring and mitigation plans.
Human Oversight	Keeping humans in final control over critical decisions, avoiding uncritical automation.	Human approval over content, scoring, flags, reports; documented checkpoints, escalation and override; auditability and right to adjust human oversight configuration.
Security & Privacy	Protecting data, content, and systems when AI/GenAI is involved.	Clear data ownership; prohibitions on sending secure data to public AI/GenAI systems; secure hosting and access controls; AI/GenAI-specific pen tests and updates.
System Maintenance & Governance	Managing model updates, data governance, and internal capacity over time.	Prior review of impactful changes; version histories and impact analyses; documented data governance; sufficient internal expertise and infrastructure.

Aspect	Focus	Vendor Expectations
Testing, Evaluation, & Monitoring	Verifying how AI/GenAI is checked before and during use in each lifecycle phase.	Defined pre-deployment checks; operational evidence deliverables; special monitoring expectations for automated scoring.
Risk Management	Developing a structured, ongoing approach to risk across governance, lifecycle, and impact.	Support for risk governance, AI/GenAI-specific testing and evaluation plan, human–AI/GenAI interaction guidance, stakeholder engagement, incident management, and supply-chain risk controls.
State Department Rights	Ensuring the agency retains control over where and how AI/GenAI is used.	Final decision-making authority on functions, stakes, and major changes; ability to require pilots, transitions, and rollbacks.

SECTION 3 - PHASE-BY-PHASE LIFECYCLE ANALYSIS

This section provides a walkthrough of possible AI/GenAI capabilities, associated risks, and suggested vendor requirements for different phases of the assessment lifecycle. Accordingly, the information in each subsection is presented in three blocks labeled "Applications", "Risks", and "Contract Levers."

The information in this section is informed, in large part, by the recent landscape scan of Lorié and Dadey (2026). We note, however, that the state of the art is rapidly evolving, which means that the specific capabilities of the systems developed or used by any given potential vendor are constantly evolving as well. Thus, opportunities to update and refine systems will become much more frequent in the future.

3.1 Construct Definition & Blueprinting

Construct definition is the systematic process of clarifying exactly what underlying knowledge, skills, abilities, or dispositions an assessment is intended to represent, drawing on empirical research in the learning sciences, syntheses of the scholarly literature, and evolving understandings of these constructs in professional practice.

Blueprint development is the process of specifying the key design dimensions of an assessment—such as the constructs covered, task or item types, and the relative weighting of sections—to ensure alignment with intended learning outcomes.

Applications

AI/GenAI systems can:

- Synthesize standards, curriculum documents, and research literature into draft construct descriptions and proposed blueprint structures for human review.
- Map existing items to construct dimensions and identify gaps, redundancies, or misalignments in the blueprint.

Risks

AI/GenAI systems may:

- Introduce construct drift by overemphasizing patterns in training data and omitting or distorting critical content elements.
- Generate synthesized descriptions that appear authoritative but contain undocumented assumptions, omissions, or inaccuracies.
- AI/GenAI-assisted proposals for constructs and blueprints may have subtle, unexpected bias and fairness implications for specific student groups or educational contexts that are not immediately obvious from quantitative technical quality evidence alone.

Contract Levers

Vendors should:

- Trace and document how AI/GenAI-generated proposals of constructs and blueprints link to human-approved source materials and how final construct and blueprint decisions were reached.
- Provide full visibility into the processes and core specifications for AI/GenAI-assisted blueprint proposals and allow the state department to require revisions or reject proposed definitions or blueprints.
- Affirm that the state department retains final approval authority over all construct definitions and blueprints.
- Include potential bias and fairness evaluations for AI/GenAI-assisted construct and blueprint options, including any subgroups that may be differentially affected, and shall document the state department's final decisions and rationale.
- Describe security precautions that prevent construct definitions - especially for emerging ones such as AI literacy - and associated blueprints from becoming publicly available.

3.2 Item & Task Development

Item and task development is the process of designing, drafting, and refining assessment prompts or questions so they elicit clear evidence of the targeted constructs and meet defined quality and fairness standards.

Applications

AI/GenAI systems can:

- Draft items, stimuli, distractors, scoring rubrics, and technology-enhanced tasks, and propose revisions to human-authored content.
- Generate systematic variants of items that vary surface features, cultural context, or format while preserving targeted construct elements.
- Conduct preliminary checks for language complexity, potential bias, accessibility issues, and basic item flaws before human review.

Risks

AI/GenAI systems may:

- Hallucinate or misrepresent disciplinary facts, relationships, or empirical data used in stimulus materials.

- Generate content that introduces subtle safety or well-being concerns (e.g., age-inappropriate scenarios or examples).
- Encode cultural, linguistic, or contextual biases that disadvantage specific student groups (e.g., multilingual learners from specific cultural or linguistic groups).
- Produce items that misrepresent or under-represent the intended construct or that introduce construct-irrelevant drivers of difficulty.
- Generate technology-enhanced tasks with incomplete, incorrect, or opaque scoring logic.
- Miss problematic items or over-flag items that are psychometrically or substantively acceptable.
- Leak or recreate secure or copyrighted content if trained or configured improperly.

Contract Levers

Vendors should:

- Use curated, licensed training corpora and prompts for content generation appropriate for educational assessment and implement controls to prevent the reproduction of copyrighted or secure materials.
- Incorporate safety and student-well-being checks into AI/GenAI-assisted item review protocols and document any AI/GenAI-related safety concerns identified, associated mitigations, and final disposition for items and stimuli.
- Require human item writers and committee members to review, revise, and approve all AI/GenAI-generated or AI/GenAI-assisted content before it enters field testing or operational use.
- Confine AI/GenAI content generation and storage to secure vendor-managed environments that meet the state department's security requirements.
- Provide documented procedures for detecting and mitigating hallucinations, construct misalignment, and bias in AI/GenAI-generated items.
- Attest, upon request, to the intellectual-property provenance of AI/GenAI-generated content and establish safeguards against reproducing protected materials.

3.3 Field Testing, Calibrating, & Equating

Field testing is the process of administering draft assessment items or tasks to representative samples of learners to gather data on how they function and to identify issues with difficulty, clarity, or fairness.

Calibrating is the process of estimating item or task parameters on a chosen measurement model so that items, tasks, and scores align on a shared scale that reflects the targeted construct.

Equating is the process of statistically linking scores from different test forms so that scores can be used interchangeably and interpreted on a common scale across administrations.

Applications

AI/Gen AI systems can:

- Generate and run code to execute psychometric analyses efficiently across multiple models and specifications.
- Predict item parameter ranges using item features and historical data to inform field-test design and sample allocation.

- Simulate response data and multi-stage testing pathways to explore design options and stress-test linking strategies.

Risks

AI/GenAI systems may:

- Produce unreliable parameter predictions when trained on limited or unrepresentative data.
- Suggest calibration or equating solutions that conflict with established standards or create unintended comparability issues.
- Influence MST routing strategies in ways that degrade measurement precision or fairness across student groups.

Contract Levers

Vendors should:

- Use AI/GenAI-generated predictions and simulations only as supplemental inputs but not as wholesale replacements for empirical field-test calibration or equating.
- Document how AI/GenAI-assisted predictions are combined with empirical field-test results when developing scaling, linking, and equating solutions.
- Include AI/GenAI-assisted design and prediction methods in the test and evaluation plan for field testing and equating, with explicit documentation of scenarios where AI/GenAI-based recommendations are overridden by psychometric judgment, along with the rationale for such decisions.
- Provide evidence that AI/GenAI-assisted design choices meet relevant psychometric standards and do not undermine trend or subgroup comparability.

3.4 Administration & Test Security

Test administration is the process of organizing and delivering an assessment under standardized conditions so that all test takers have a consistent and fair testing experience.

Test security is the set of policies and procedures used to protect assessment content, safeguard test-taker data, and prevent cheating or unauthorized access that could compromise score validity.

Applications

AI/Gen AI systems can:

- Monitor response patterns and ancillary data to flag potential irregularities, impersonation, or device anomalies.
- Optimize scheduling, form assignment, and device utilization to support smooth operational administration.
- Deliver adaptive accessibility supports where appropriate and pre-approved, such as controlled vocabulary supports or presentation adjustments.

Risks

AI/GenAI systems may:

- Disproportionately flag specific groups of students (e.g., English learners, students with disabilities) as high risk, resulting in disparate impact.
- Use data sources or monitoring approaches that exceed acceptable privacy or surveillance expectations.

- Lead to inappropriate actions against students or schools, particularly for vulnerable groups, due to an over-reliance on AI/GenAI-based monitoring or anomaly detection.
- Alter the construct being measured when dynamically generating or adapting accessibility supports.
- Consume computational resources at a scale that compromises test delivery performance.

Contract Levers

Vendors should:

- Conduct and document fairness and accuracy evaluations of AI/GenAI-based security and monitoring systems, including subgroup analyses.
- Ensure that any flags or high-stakes determinations generated by AI/GenAI systems undergo human review before sanctions or security actions are imposed.
- Implement and periodically review decision rules, training, and supporting materials for human reviewers who interpret AI/GenAI-generated security flags, with a focus on minimizing unjustified adverse impacts on students and schools.
- Validate that AI/GenAI-driven accessibility supports do not compromise construct validity or produce unintended score differences.
- Prohibit the use of unapproved real-time public AI services for any aspect of test administration or security monitoring.
- Demonstrate that infrastructure supporting AI/GenAI functions can reliably handle peak testing loads without degrading student experience.

3.5 Response Scoring

Response scoring is the process of evaluating test-taker answers against predefined scoring rules or rubrics to produce consistent, interpretable scores that reflect the targeted constructs.

Applications

AI/GenAI systems can:

- Score constructed responses, essays, and multimodal responses in alignment with human-developed rubrics.
- Identify ambiguous rubric language and propose refinements, as well as highlight key response segments for human scorers.
- Monitor human scoring patterns for potential drift and suggest targeted quality-control reviews.

Risks

AI/GenAI systems may:

- Produce biased scores for specific student groups, language varieties, or response styles.
- Yield inconsistent scores for substantively identical responses under slightly varying conditions.
- Elevate exposure risk if scoring pipelines or model endpoints are not properly secured.

Contract Levers

Vendors should:

- Configure AI/GenAI scoring models to produce replicable scores under defined operational conditions.

- Perform and report regular bias, fairness, reliability, and drift analyses for AI/GenAI-assisted scoring, including disaggregated subgroup results.
- Implement clear routing rules to send uncertain or out-of-distribution responses to human scorers and document these rules.
- Provide scoring documentation that describes model behavior in edge cases, subgroup performance, and thresholds for diverting responses to human scoring, together with residual-risk statements for each operational scoring model.
- Obtain the state department approval for any foundational change to scoring models and provide sufficient documentation for third-party replication and review.

3.6 Reporting & Interpretation

Reporting is the process of summarizing and communicating assessment results in clear, user-friendly formats that highlight key findings and support educational decision-making.

Interpretation is the process of making sense of assessment results by drawing valid inferences about what scores mean in relation to the targeted constructs, learning goals, and appropriate uses.

Applications

AI/GenAI systems can:

- Generate general draft interpretive narratives, FAQs, and training materials tailored to different stakeholder audiences and languages.
- Generate custom visualizations or tabulations of information based on user requests.
- Generate custom explanations of output for users with different levels of technical understanding, trained only on relevant content to avoid hallucinations.

Risks

AI/GenAI systems may:

- Recommend uses of scores or instructional actions that exceed validated purposes.
- Hallucinate specific score values, growth trajectories or error estimates and misrepresent trends.
- Introduce inconsistencies across language versions or reporting modalities for resources.
- Create narratives or visualizations that are persuasive yet misleading, increasing the risk that educators or families misinterpret the meaning or precision of scores.

Contract Levers

Vendors should:

- Validate narrative templates, generation procedures, and translation processes prior to operational use, including stress-testing for accuracy and consistency.
- Provide the state department with access to report-generation logic, templates, and representative outputs for review and audit.
- Clearly distinguish AI/GenAI-generated draft materials from finalized, the state department-approved reporting resources.
- Evaluate AI/GenAI-generated reporting materials with representative user groups (such as educators, administrators, and families) and shall revise templates and guardrails where users misinterpret the meaning, limits, or intended uses of scores.

Table 3 on the following pages summarizes the key ideas from this section one more time in a compact form. This concludes the section on applications, risks, and contract levers in this brief. In the appendix that follows we now provide sample contractual language that translates these considerations into explicit directive statements.

Table 3. Summary of Phase-by-Phase Lifecycle Analysis

Phase	Illustrative AI/GenAI Applications	Major Risks	Key Contract Levers
Construct Definition & Blueprinting	Summarizing standards and research into draft constructs; proposing blueprint structures; mapping items to construct dimensions for human review.	Construct drift; over-confident synthesized descriptions; subtle fairness issues across student groups.	Require traceability to source materials; ensure state approval of final constructs and blueprints; include bias and fairness checks; protect emerging constructs (e.g., AI/GenAI literacy) from disclosure.
Item & Task Development	Drafting items, stimuli, distractors, rubrics, and technology-enhanced tasks; generating item variants; pre-screening content for language complexity, bias, and accessibility issues.	Hallucinated or inaccurate content; safety or well-being concerns; cultural and linguistic bias; construct misalignment; incomplete or opaque scoring logic; leakage of secure or copyrighted content.	Use curated, licensed training data; embed safety and bias review; require human review and approval; confine generation to secure environments; document hallucination, misalignment, and bias detection procedures; safeguard content provenance.
Field Testing, Calibrating, & Equating	Predicting item parameter ranges; simulating response data and multi-stage designs; generating analysis code to support psychometric studies.	Unreliable predictions from limited data; recommendations that conflict with measurement standards; degraded precision or fairness; unintended comparability issues across administrations.	Use AI/GenAI predictions only as a supplement to empirical data; document how predictions and field-test results are combined; require that AI/GenAI-assisted choices meet psychometric standards and protect trend and subgroup comparability.

Phase	Illustrative AI/GenAI Applications	Major Risks	Key Contract Levers
Administration & Test Security	Monitoring response patterns and ancillary data for irregularities, impersonation, or device anomalies; optimizing scheduling and form assignment; supporting controlled, pre-approved accessibility adjustments.	Disproportionate flags for specific groups; privacy or surveillance concerns; over-reliance on AI/ GenAI-based monitoring; construct shifts from dynamic supports; strain on delivery infrastructure.	Conduct fairness and accuracy evaluations of monitoring systems; require human review before sanctions; constrain data sources and monitoring practices; periodically review decision rules and operational impacts.
Response Scoring	Automating or assisting scoring processes; monitoring human scorer performance; flagging anomalous patterns for review.	Biased or unstable scores; drift across model updates; opaque scoring logic; over-trust in AI/ GenAI-generated scores.	Define readiness and monitoring criteria; require subgroup analyses; maintain human oversight and clear escalation thresholds; document methods and evidence in technical and peer-review materials.
Reporting & Interpretation	Drafting reports, explanations, and exemplars; tailoring narratives and summaries for different stakeholder audiences.	Misleading or hallucinated interpretations; inconsistent alignment with policy; inequitable or confusing messaging for different groups.	Require human review and approval of reporting content; label AI/ GenAI-generated or assisted content; provide guidance on appropriate use; define incident response for reporting errors or misinterpretations.

APPENDIX: ILLUSTRATIVE CONTRACT LANGUAGE

The following clauses are examples illustrating expectations; for appropriate legal language, state departments should consult an attorney.

There are often no clear-cut categories with which a particular phrase can be associated so it is best to think of the boldfaced headers as "primary tags" that help organize the information, rather than as "singular, definitive" categories.

Some of the statements are also variations of one another that foreground different aspects to help reflect on what aspects are most important.

Moreover, working explicitly on language across assessment, content, and procurement teams, among others, can help teams:

- discuss, reconcile, and document the core values and guardrails that matter for their work
- reconcile different levels of desired grain size (e.g., one principle for all phases vs. separate phase-specific variants)
- identify which concepts are of strategic priority (e.g., accuracy or consistency vs. validity or fairness)
- agree on specific time frames (e.g., describing how a system works now vs. how it is expected to evolve)

Importantly, they should be drafted early enough before key RFP release deadlines to allow for thoughtful reconciliation and refinement. They should be reviewed both individually as well as holistically across an RFP to ensure coherence, coverage, and clarity.

For example, it can be helpful to use an AI/GenAI tool to extract all specific statements that signal vendor expectations into a separate document and review these for potential conflicting signals, missed considerations, and unnecessary redundancies.

Section 2

Transparency & Explainability (Section 2.1)

- "The vendor shall maintain and provide to the state department, upon request and at least annually, a complete inventory of all AI/GenAI components used in any assessment process, including each component's purpose, inputs, outputs, and placement in the assessment workflow."
- "The vendor shall maintain and provide to the state department, upon request and at least annually, an inventory of all third-party AI/GenAI components, services, libraries, and application programming interfaces used in any assessment-related AI/GenAI system, including their provider, version, intended use, and known limitations."
- "The vendor shall disclose to the state department, in writing, the material ways in which each AI/GenAI component affects student scores, interpretations, test security, process efficiency, and reporting, including known limitations and residual risks."
- "The vendor shall provide the state department with documentation describing the nature and categories of training and fine-tuning data, evaluation metrics, and validation results for each AI/GenAI system used under this Contract."

- "The vendor shall develop and maintain plain-language explanations, FAQs, and technical summaries that the state department may share with districts, educators, students, families, and the public to explain where and how AI/GenAI is used within assessment."
- "Upon request, the vendor shall provide the state department with de-identified examples of AI/GenAI-assisted content, scoring outputs, and reports that the state department may use in stakeholder engagement and public communication, subject to applicable security and confidentiality requirements."
- "For each material AI/GenAI component used in assessment, the vendor shall provide the state department with documentation that explains, at an appropriate level of technical detail, how the system transforms inputs into outputs, including key features used, major model families or architectures, and the factors that most strongly influence decisions or scores, in a form that can be understood by technically informed the state department staff and external reviewers."
- "Where technically feasible, the vendor shall implement and document mechanisms that support case-level explainability for AI/GenAI-assisted decisions (e.g., scoring rationales, reasons for security flags, or basis for automated alerts), so that the state department and its designees can understand and, when needed, challenge or override specific AI/GenAI outputs without compromising test security or proprietary information."

Bias & Fairness (Section 2.2)

- "The vendor shall adopt and document a bias and fairness assessment framework for all AI/GenAI systems used under this Contract, including definitions of fairness, student groups examined, metrics employed, and decision criteria for acceptable performance."
- "The vendor shall ensure that bias and fairness evaluations cover all material AI/GenAI uses in assessment, including construct and blueprint development, item and task generation, field-test design, administration and security monitoring, scoring, and reporting, with phase-specific metrics and thresholds defined in consultation with the state department."
- "Prior to operational deployment and at each major model update, the vendor shall conduct bias and fairness analyses for AI/GenAI systems and provide the state department with summary results, including subgroup-disaggregated metrics and mitigation actions taken where necessary."
- "If any AI/GenAI system exhibits disparities that exceed agreed-upon thresholds, the vendor shall notify the state department within X business days and implement a remediation plan acceptable to the state department, which may include disabling or constraining use of the affected system."
- "The vendor shall not deploy or continue to operate an AI/GenAI system in a high-stakes context if bias and fairness evidence indicates that it introduces material, unmitigated disparities for legally protected or the state department-priority student groups, unless the state department explicitly authorizes constrained use with documented safeguards."
- "When AI/GenAI tools are used to identify potentially biased items, tasks, or reporting narratives, the vendor shall document how human reviewers considered AI/GenAI-generated flags, including the disposition of flagged content and rationales for retaining, revising, or removing it."
- "The vendor shall periodically review and, as needed, update its bias and fairness framework, metrics, and thresholds for AI/GenAI systems in light of new research, standards, or the state department policy changes, and shall discuss proposed updates with the state department prior to adoption."

- "The vendor shall periodically review emerging research and tools for mitigating bias in AI/GenAI-assisted assessment processes (e.g., new fairness metrics, debiasing algorithms, or data-augmentation strategies) and, in consultation with the state department, shall evaluate whether any such approaches are appropriate for assessment; where the state department approves their use, the vendor shall pilot and document their effects on validity, fairness, and operational feasibility before broader adoption."

Human-in-the-Loop Oversight (Section 2.3)

- "The vendor shall ensure that human experts, including the state department-designated committees, retain final approval authority for all assessment content, scoring methodologies, and security determinations influenced by AI/GenAI systems."
- "AI/GenAI-generated or AI/GenAI-assisted recommendations affecting scoring, routing, or test security shall not be used as the sole basis for any high-stakes decision and shall be subject to documented human review."
- "The vendor shall maintain audit logs of AI/GenAI-influenced decisions, including inputs, outputs, human review actions, and final determinations, and shall provide such logs to the state department upon request."
- "The vendor shall define, in consultation with the state department, decision thresholds and contexts in which AI/GenAI-generated recommendations must be escalated for human review and shall ensure that no AI/GenAI output alone triggers high-stakes actions affecting students, educators, or schools."
- "The vendor shall provide training materials and, when requested, training sessions for the state department staff and other designated users that explain: (a) where AI/GenAI is used in assessment processes, (b) how to interpret AI/GenAI-generated outputs such as scores, flags, or reports, and (c) how to exercise human judgment in reviewing and overriding AI/GenAI recommendations."
- "The vendor shall implement user-interface and workflow designs that clearly distinguish AI/GenAI-generated content from human-authored content, indicate when AI/GenAI has influenced a recommendation, and make it straightforward for human reviewers to override or request reconsideration of AI/GenAI suggestions."
- "The vendor shall periodically evaluate how human reviewers interact with AI/GenAI-assisted tools (e.g., rates of overrides, agreement patterns, and response to flags) and, in consultation with the state department, shall adjust training, thresholds, or workflows where evidence suggests over-reliance on or systematic disregard of AI/GenAI outputs."
- "In any pilot or phased rollout of new AI/GenAI-assisted processes, the vendor shall maintain enhanced human-in-the-loop checkpoints and shall not remove or relax those checkpoints in operational use without the state department's explicit written approval based on supporting evidence."
- "The vendor shall ensure that workflows and performance-management practices for staff who review AI/GenAI-assisted outputs (including scorers, security reviewers, and the state department personnel) encourage appropriate use of professional judgment, including overriding AI/GenAI recommendations where warranted, and shall not penalize staff solely for reasonable, documented disagreements with AI/GenAI-generated suggestions."

Security & Privacy (Section 2.4)

- "The vendor shall not transmit or expose student data, secure content, or scoring materials to external AI/GenAI services, public APIs, or third-party systems that have not been explicitly approved in writing by the state department."
- "All AI/GenAI training, fine-tuning, and inference activities involving state department data shall occur within secure environments that meet or exceed the security requirements specified in the RFP and this Contract, including encryption, access controls, logging, and monitoring."
- "The vendor shall perform AI/GenAI-specific penetration testing and adversarial ('red-team') exercises at least annually and after major system changes and shall provide the state department with a high-level summary of findings and remediation steps."
- "Any use of state department data for training or fine-tuning AI/GenAI models under this Contract shall be limited to models dedicated to assessment or to other the state department-authorized purposes, and shall be accompanied by documentation describing data selection, preprocessing, safeguards against memorization of identifiable student content, and mechanisms to prevent reidentification."
- "Upon termination or expiration of this Contract, the vendor shall, at the state department's direction, delete, return, or irreversibly de-identify all state department data used in AI/GenAI training, fine-tuning, or evaluation, and shall provide written certification of completion, subject to legally required retention obligations."
- "The vendor shall not use the state department assessment data, including student responses, scores, metadata, or secure content, to train, fine-tune, or otherwise improve general-purpose commercial AI/GenAI models or products offered to other customers, unless explicitly authorized in writing by the state department."
- "The vendor shall ensure that agreements with third-party AI/GenAI providers include provisions that are at least as protective of the state department's data, security, privacy, and intellectual-property rights as this Contract, and shall be responsible for any breaches or noncompliance by such third parties in connection with assessment activities."
- "The vendor shall apply data-minimization principles to all AI/GenAI-assisted processes, collecting, accessing, and retaining only those student-, educator-, and school-level data elements that are strictly necessary to perform contractually authorized functions, and shall document how such determinations are made."
- "At the state department's request, the vendor shall provide a data-flow and data-protection description for AI/GenAI-assisted processes, including where state department data are stored and processed, how access is controlled and logged, and how privacy protections are enforced for students in small schools or sparse subgroups that may be at higher risk of reidentification."

System Maintenance & Governance (Section 2.5)

- "For each contract year, the vendor shall provide the state department with a concise AI/GenAI system governance report summarizing key changes to AI/GenAI components, major incidents or issues, mitigations implemented, and plans for upcoming improvements that may affect assessment."
- "The vendor shall maintain a version-controlled record for each AI/GenAI component, including release dates, change descriptions, validation evidence, and any known impacts on longitudinal comparability, and shall provide such records to the state department upon request."

- "The vendor shall maintain and follow written data governance policies covering data quality, lineage, retention, access, and disposal for all data used by AI/GenAI systems under this Contract."
- "The vendor shall maintain a formal lifecycle register for AI/GenAI components used in assessment, including dates of initial deployment, major updates, risk level, and scheduled review points, and shall provide the state department with a summary of this register at least annually."
- "For any AI/GenAI component that materially affects scoring, routing, reporting, or test security, the vendor shall document decommissioning and rollback procedures, including strategies to protect longitudinal score comparability, and shall provide such documentation to the state department upon request."
- "If an AI/GenAI component must be decommissioned or replaced, the vendor shall develop, in consultation with the state department, a succession plan that describes transition timelines, any necessary parallel-run periods, impact on reporting, and communications to stakeholders."
- "The vendor shall support the state department's efforts to communicate with and gather feedback from districts, educators, and other stakeholders about the use of AI/GenAI in assessment by providing, upon request, explanatory materials, sample outputs, and staff participation in the state department-led meetings or trainings."
- "Within X calendar days of deploying a new AI/GenAI component or materially modifying an existing component, the vendor shall notify the state department and provide updated documentation, including high-level model descriptions, evaluation methods, and hosting arrangements."
- "The vendor shall ensure that internal governance bodies overseeing AI/GenAI use (e.g., AI/GenAI review boards or change-control committees) consider assessment-specific validity, fairness, and policy requirements when reviewing proposed changes, and shall make relevant portions of their decisions available to the state department upon request."

Phase-wide Test, Evaluation, and Monitoring (Section 2.6)

- "The vendor shall develop and maintain a written AI/GenAI test, evaluation, and monitoring plan that covers all phases of the assessment lifecycle in which AI/GenAI systems or AI/GenAI-assisted processes are used, including construct and blueprint development, item and task development, field testing, calibrating, and equating, administration and security, scoring, and reporting. The plan shall describe pre-deployment checks, in-operation monitoring procedures, metrics, and acceptance thresholds for each relevant phase."
- "For each material AI/GenAI component, the vendor shall define and document pre-deployment evaluation procedures, including test datasets, performance metrics, subgroup analyses where appropriate, and criteria for determining whether the component is fit for its intended use in the relevant phase."
- "The vendor shall implement ongoing monitoring for AI/GenAI components used in operational assessment activities, including periodic checks for performance, bias, drift, and stability. The vendor shall summarize monitoring results for the state department at least annually, or more frequently upon request, and shall identify any required mitigations or configuration changes."

- "When monitoring reveals material degradation in performance, fairness, or stability of an AI/GenAI component, the vendor shall notify the state department, propose remediation steps (such as retraining, reconfiguration, or temporary suspension), and implement mutually agreed changes within timelines acceptable to the state department."
- "The vendor shall provide the state department with documentation sufficient to support independent review of AI/GenAI test, evaluation, and monitoring activities, including descriptions of datasets, metrics, thresholds, and any changes made over time, in a form that the state department can incorporate into technical reports and oversight documentation."
- "At intervals agreed with the state department, and at least once per contract term, the vendor shall prepare an AI/GenAI impact-assessment summary describing how AI/GenAI components used in assessment have performed with respect to validity, fairness, security, privacy, and educational impact, including relevant subgroup analyses where available."
- "At the state department's request, the vendor shall provide comparative evidence showing performance, fairness, and stability of any replacement AI/GenAI component relative to the prior component, with particular attention to impacts on trend reporting and subgroup comparability."
- "The vendor shall obtain prior written approval from the state department before implementing any foundational change to AI/GenAI systems that materially affects scoring, routing, reporting, or test security."

Risk Management (Section 2.7)

- "The vendor shall develop, maintain, and provide to the state department an AI/GenAI risk-management plan covering all AI/GenAI systems and AI/GenAI-assisted processes used in any assessment activity. The plan shall describe major AI/GenAI components, intended uses, key risks, safeguards, and monitoring procedures and shall be updated at least annually and after any material change to AI/GenAI components."
- "For each material AI/GenAI component, the vendor shall document known limitations and residual risks, including potential impacts on validity, fairness, security, privacy, and score interpretation, and shall provide the state department with updated residual-risk summaries following any major model update or configuration change."
- "The vendor shall establish and implement AI/GenAI-specific incident-management procedures that cover detection, classification, escalation, remediation, and documentation of AI/GenAI-related incidents. AI/GenAI-related incidents that may materially affect scores, classifications, reporting, or test security shall be reported to the state department in writing within the timelines specified by the state department, together with a description of scope, suspected causes, and interim mitigations."
- "The vendor shall work with the state department to identify any material concerns raised through stakeholder feedback or impact-assessment results and shall propose and implement mutually agreed adjustments to AI/GenAI configurations, processes, or safeguards to address those concerns."
- "The vendor shall ensure that AI/GenAI-assisted tools used by the state department staff, scorers, educators, or other designated users are accompanied by clear guidance on appropriate and inappropriate uses of AI/GenAI outputs, including examples of when human review, escalation, or override is required."

- "The vendor shall maintain documentation describing decision rules, thresholds, and routing logic that govern when AI/GenAI outputs are accepted automatically, when they are flagged for human review, and when they are overridden, and shall provide such documentation to the state department upon request."
- "The vendor shall ensure that AI/GenAI-assisted recommendations affecting high-stakes decisions are never used as the sole basis for such decisions and that final authority rests with human decision-makers designated by the state department, as described in the state department's policies and procedures."
- "The vendor shall maintain contingency and business-continuity plans for critical third-party AI/GenAI components whose behavior, availability, licensing, or terms of use may change, to ensure uninterrupted assessment operations and protection of validity, fairness, and security."

State Department Rights (Section 2.8)

- "The state department retains final authority to determine whether AI/GenAI systems may be used for any given assessment function, including but not limited to scoring, test security, routing, and reporting, and the vendor shall implement AI/GenAI components only in functions and contexts explicitly authorized by the state department."
- "The vendor shall not move any AI/GenAI component from pilot use or low-stakes use into fully operational, high-stakes use without the state department's prior written approval and agreement on any conditions, limitations, or additional safeguards associated with that use."
- "The vendor shall obtain the state department's prior written approval for any foundational change to AI/GenAI components that materially affects scoring, routing, reporting, or test security, including changes to model architecture, training or fine-tuning data, operating thresholds, or deployment environments."
- "Upon request, the vendor shall provide the state department with impact analyses, validation evidence, and implementation plans for proposed foundational changes to AI/GenAI components, including any proposed transition periods, parallel-run strategies, or back-out procedures needed to protect longitudinal comparability and operational continuity."
- "If the state department determines that continued use of a particular AI/GenAI component is no longer acceptable for policy, technical, or risk reasons, the vendor shall cooperate with the state department to suspend, decommission, or replace that component on a mutually agreed schedule and shall support any necessary transition arrangements identified by the state department."
- "Before introducing or materially changing any critical third-party AI/GenAI component that could affect scoring, routing, reporting, or test security, the vendor shall notify the state department in writing, provide relevant risk and impact analyses, and obtain the state department's written approval."
- "All student data, secure test content, scoring information, and related assessment artifacts remain the exclusive property of the state department; the vendor is granted only a limited, non-exclusive license to use such data to perform obligations under this Contract."
- "The vendor shall not deploy or materially modify any AI/GenAI component that materially affects scoring, routing, reporting, or test security into operational use without prior written notice to the state department and an opportunity for the state department to review relevant documentation from the AI/GenAI risk-management plan."

Section 3

Construct & Blueprint Development (Section 3.1)

- "Any AI/GenAI-generated proposals for construct definitions, claims, targets, or blueprints shall be treated as draft inputs; the state department retains sole authority to approve or reject final constructs and blueprints."
- "The vendor shall provide documentation that traces AI/GenAI-generated blueprint proposals to the underlying standards, curriculum frameworks, and reference materials used as inputs, together with a narrative of how human judgment was applied."
- "The vendor shall document, for any AI/GenAI-assisted construct or blueprint proposals, how proposed domains, claims, and reporting categories map to the state department-approved standards, curriculum documents, and policy priorities, including a clear indication of which elements were AI/GenAI-generated versus human-authored."
- "Prior to the state department approval of any AI/GenAI-assisted construct or blueprint revisions, the vendor shall provide an analysis of potential impacts on different student groups (e.g., by disability status, English-learner status, race/ethnicity, or program participation), including any identified risks to equity of opportunity to learn or fairness of score interpretations."
- "The vendor shall ensure that AI/GenAI-generated construct or blueprint options are treated as advisory inputs only and shall not be implemented operationally unless a human committee designated by the state department has reviewed alternative formulations, documented its rationale, and explicitly approved the selected configuration."
- "The vendor shall maintain version-controlled records of all construct and blueprint proposals for assessment, including AI/GenAI-assisted drafts, human edits, rejected alternatives, and final approved versions, and shall provide such records to the state department upon request to support audit, peer review, and longitudinal comparability studies."
- "When AI/GenAI is used to surface gaps, redundancies, or misalignments in the blueprint, the vendor shall document how those AI/GenAI-identified issues were evaluated by human experts and whether they led to changes in the construct or blueprint, including justification for accepting or rejecting AI/GenAI-generated recommendations."

Item & Task Development (Section 3.2)

- "The vendor shall ensure that all AI/GenAI-generated or AI/GenAI-assisted items, stimuli, distractors, and rubrics undergo human review and approval by qualified content experts and the state department-designated committees prior to use in field testing or operational forms."
- "The vendor shall ensure that AI/GenAI-assisted item and task development supports, including automated checks for language complexity, cultural references, and accessibility, are treated as advisory tools and do not replace established human review processes required by the state department."
- "The vendor shall not use general-purpose public internet content or unvetted web crawls as the sole training source for AI/GenAI systems generating test content and shall implement safeguards to prevent reproduction of copyrighted or secure materials."
- "AI/GenAI-generated items shall be clearly identifiable within the vendor's authoring and review systems to support tracking, review, and audit."
- "Upon request, the vendor shall provide the state department with descriptions of prompts, guardrails, and filters used to constrain AI/GenAI-generated content for assessment use."

- "The vendor shall implement documented procedures for detecting and mitigating AI/GenAI-related issues such as factual inaccuracies, hallucinated content, construct-irrelevant difficulty, and subtle safety or well-being concerns in AI/GenAI-generated items and tasks and shall make such procedures available to the state department upon request."
- "When AI/GenAI tools are used to generate systematic item variants or parallel forms, the vendor shall provide evidence that variants preserve the intended construct coverage and difficulty range and shall monitor for unintended differential functioning across student subgroups."
- "If AI/GenAI-assisted item review tools are used to flag potentially problematic content (e.g., biased, offensive, or construct-misaligned items), the vendor shall document the nature of flagged content, including reasons for accepting, revising, or rejecting items, and shall provide the state department with summaries upon request."
- "The vendor shall document and evaluate scoring rules and AI/GenAI-assisted scoring approaches for complex, multi-action tasks to ensure that partial credit, alternative solution paths, and common misconceptions are handled in ways that are consistent with the state department-approved scoring policies and rubrics."

Field Testing, Calibrating, & Equating (Section 3.3)

- "The vendor shall not rely solely on AI/GenAI-predicted item parameters, scale relationships, or linking functions for operational use; empirical field-test and calibration data shall remain the primary basis for scaling decisions."
- "When AI/GenAI-assisted predictions or simulations inform field-test designs or equating plans, the vendor shall document the role of AI/GenAI, the underlying assumptions, and the validation evidence supporting those uses, and shall make such documentation available to the state department."
- "The vendor shall ensure that AI/GenAI-assisted predictions of item parameters or routing behavior are used only to inform design decisions and shall not replace empirical field-test data in determining final calibration, scaling, or equating solutions."
- "For any field-test or equating design that relies on AI/GenAI-generated simulations or predictions, the vendor shall provide the state department with documentation describing the input data, modeling assumptions, performance of the predictions against observed data, and rationale for accepting or rejecting AI/GenAI-generated recommendations."
- "The vendor shall evaluate the impact of AI/GenAI-assisted design choices on measurement precision and comparability for key student subgroups and shall not implement AI/GenAI-driven design changes that materially degrade subgroup comparability or trend interpretation without the state department's explicit written approval."
- "When AI/GenAI tools are used to generate code or specifications for calibration and equating analyses, the vendor shall subject such code to standard quality-assurance and independent review processes and shall retain human responsibility for final methodological decisions and documentation."
- "The vendor shall maintain a version-controlled record of AI/GenAI-assisted design experiments (e.g., alternative routing or linking strategies suggested by AI/GenAI) and their evaluation outcomes and shall provide summaries to the state department upon request to support audit and peer-review needs."

- "When AI/GenAI tools propose alternative IRT models or equating approaches (e.g., different linking chains, anchor designs, or item inclusion rules), the vendor shall document the alternatives considered, the evaluation criteria applied, and the rationale for the final selection, including reasons for accepting or rejecting AI/GenAI-generated options."
- "Any code, scripts, or simulation studies generated or materially assisted by AI/GenAI for field testing, IRT calibration, or equating shall undergo standard quality-assurance procedures, including independent review or replication by qualified staff, with documentation of checks performed and issues resolved."

Administration & Test Security (Section 3.4)

- "The vendor shall validate AI/GenAI-based security and anomaly-detection tools, including evaluation of false-positive and false-negative rates by student subgroup, and shall share summary results with the state department."
- "Any AI/GenAI-generated security flags or risk scores shall be treated as preliminary indicators; human decision-makers shall make final determinations regarding invalidations, investigations, or sanctions under procedures approved by the state department."
- "The vendor shall ensure that no AI/GenAI-based monitoring practice implemented under this Contract conflicts with applicable privacy laws, state policy, or the state department-approved test administration protocols."
- "The vendor shall demonstrate, upon request, that AI/GenAI-enabled services used during test administration are capable of supporting peak loads without compromising test delivery performance."
- "The vendor shall ensure that AI/GenAI-enabled accessibility features, including adaptive presentation, language supports, and assistive technologies, comply with applicable federal and state disability and civil-rights laws and the state department accessibility policies, and shall document how these features preserve the intended construct."
- "The vendor shall conduct and provide to the state department periodic evaluations of AI/GenAI-enabled accessibility supports by relevant student subgroups, including students with disabilities and English learners, including evidence that such supports do not introduce construct-irrelevant variance or unfair score differences."
- "The vendor shall not deploy any AI/GenAI-based monitoring or proctoring functions that rely on protected characteristics or proxies for protected characteristics and shall implement controls to minimize unjustified disparate impact on any protected group."
- "The vendor shall ensure that any data sources used by AI/GenAI-enabled administration and security tools (e.g., keystroke patterns, timing information, or device telemetry) are limited to those explicitly approved by the state department and are collected and used in a manner consistent with applicable privacy and civil-rights laws and the state department policies."
- "Prior to deploying new or substantially revised AI/GenAI-assisted administration or security tools, the vendor shall conduct pilot testing or dry runs under conditions agreed upon with the state department, evaluate operational impacts (including test-taker experience and administrator workload), and obtain the state department's written approval for operational use."

Scoring (Section 3.5)

- "AI/GenAI-based scoring engines shall be configured to produce stable, replicable scores under defined operational conditions, and any remaining sources of randomness shall be documented and controlled."
- "The vendor shall perform periodic reliability, bias, and drift analyses of AI/GenAI-assisted scoring engines, including subgroup-disaggregated results, and shall provide summary reports to the state department at least annually."
- "The vendor shall maintain documented routing rules specifying when responses are scored by AI, by humans, or by a hybrid process, and shall review and adjust these rules based on empirical evidence and the state department feedback."
- "The state department must approve any foundational change in scoring model architecture, training data, or operating thresholds before such change is used operationally."
- "The vendor shall support the state department's score-review and appeals processes by providing, upon request and subject to security constraints, information sufficient for the state department to investigate potential AI/GenAI-related scoring anomalies, including logs of routing decisions and summaries of model behavior for the responses at issue."
- "The vendor shall ensure that AI/GenAI-assisted scoring tools are configured so that the state department-approved human-developed rubrics and scoring rules remain the authoritative reference, and that any AI/GenAI-generated scoring suggestions or rationales are consistent with those rubrics."
- "When AI/GenAI tools are used to monitor human scoring quality (e.g., by flagging potential drift, inconsistencies, or anomalous patterns), the vendor shall document the rules and thresholds used, provide the state department with periodic summaries of findings, and ensure that any high-stakes scorer actions (such as retraining or decertification) are based on documented human review."
- "The vendor shall support the state department's score-review and appeals processes by providing, subject to applicable security and confidentiality requirements, information sufficient to investigate potential AI/GenAI-related scoring anomalies, including logs of scoring model versions, routing decisions, and summaries of model behavior for the responses at issue."
- "For any substantial change to AI/GenAI-assisted scoring models or workflows that could affect score distributions, subgroup relationships, or longitudinal comparability, the vendor shall provide the state department with impact analyses and shall not implement such changes operationally without the state department's explicit written approval."

Reporting & Interpretation (Section 3.6)

- "Upon the state department's request, the vendor shall provide plain-language descriptions that the state department may use to explain to educators, families, and students where AI/GenAI is used in scoring and how human review and oversight are incorporated, including general explanations of safeguards for fairness and accuracy."
- "The vendor shall validate AI/GenAI-generated interpretive narratives, FAQs, visualizations, and training materials prior to operational use, including checks for factual accuracy, alignment with validated score uses, and consistency across languages."
- "AI/GenAI-generated reporting materials shall be clearly identified as drafts until reviewed and approved by the state department or its designees, and the state department retains final authority over all published reporting content."

- "AI/GenAI-assisted analyses that model potential cut scores or reporting structures shall be treated as advisory; final standard-setting and reporting decisions shall be made using the state department-approved processes."
- "For AI/GenAI-assisted tools that present information directly to scorers, educators, or the state department staff (such as dashboards, anomaly-detection interfaces, or AI/GenAI-generated draft reports), the vendor shall conduct usability and interpretability reviews and shall incorporate the state department feedback to reduce the risk of misinterpretation or over-reliance on AI/GenAI outputs."
- "The vendor shall ensure that AI/GenAI-generated draft interpretations, recommendations, or visualizations do not expand the validated purposes of assessment scores and shall include guardrails to prevent suggestions of unvalidated high-stakes uses (e.g., individual teacher evaluation or retention decisions)."
- "For any AI/GenAI-assisted reporting that is provided in multiple languages or formats, the vendor shall document and periodically evaluate procedures to ensure that meaning, cautions, and limitations are substantively equivalent across languages and modalities, and shall provide the state department with evidence of such evaluations upon request."
- "Before implementing substantial changes to AI/GenAI-assisted reporting logic, templates, or personalization features that could affect how users interpret scores or trends, the vendor shall notify the state department, provide impact analyses (including user-testing summaries), and obtain the state department's written approval."
- "Upon the state department's request, the vendor shall supply de-identified example reports and explanatory materials, including those that incorporate AI/GenAI-generated content, that the state department may use in training, guidance, and stakeholder-engagement activities."

National Center for the Improvement
of Educational Assessment, Inc.

Dover, New Hampshire

www.nciea.org