JOHNS HOPKINS SCHOOL of EDUCATION | Institute for Education Policy

Revolutionizing ELA Assessment: Harnessing AI for Curriculum-Aligned, Evidence-Based, and Equitable Evaluation

Dr. David Steiner

Background and Problem Statement

For decades, state English Language Arts (ELA) assessments have been built on a simple but flawed premise: that reading comprehension is best measured by presenting students with previously unseen passages and asking them to demonstrate decontextualized "skills" such as "find the main idea" or "identify the author's purpose."

Cognitive science and classroom experience point in another direction. A robust body of research shows that reading comprehension is heavily dependent on **background knowledge and vocabulary**, not on free-floating skills. E.D. Hirsch, Daniel Willingham, Recht and Leslie, and others have documented that when decoding is held constant, students with stronger topic knowledge substantially outperform peers on comprehension tasks—even when those peers have comparable or better generic "reading skill." In Recht and Leslie's classic baseball study, struggling decoders who knew a lot about baseball comprehended a baseball passage better than strong decoders who lacked that knowledge. (See Appendix C for a list of resources.)

This has profound implications for assessment. When state ELA tests rely on arbitrary, unfamiliar passages, they inevitably reward **out-of-school knowledge**—books in the home, museum visits, enriched adult conversation—rather than what schools teach. Students from more advantaged backgrounds are more likely to have the broad knowledge base these tests presuppose. Students whose families and communities have been denied those resources are penalized, even when they have worked hard and engaged fully with their school curriculum.

At the same time, curriculum-agnostic ELA tests are often **instructionally unhelpful**. Because teachers cannot know which passages will appear on the state exam, they are incentivized to rely on skills-drilling and detached "test prep" that bear little resemblance to the rich, knowledge-building instruction we want in classrooms. This approach is misaligned, both with contemporary reading science and also with what parents and educators recognize as meaningful learning.

Recent (2024) NAEP reading results confirm that our current ways of teaching and testing reading aren't working. While every student demographic except the top 10% performed worse than at any time since the 1990s, economically disadvantaged students showed the greatest declines, thus widening pre-existing gaps.

In short, the prevailing model of ELA assessment is:

- Misaligned with how reading competency grows.
- Inequitable in its reliance on out-of-school knowledge.
- Weakly connected to the high-quality curricula that districts are increasingly adopting.
- Reinforcing of poor pedagogical practices.

A Better Alternative: Curriculum-Based ELA Assessments

An alternative is to build ELA assessments **directly from the content that students encounter** in well-designed, knowledge-rich curricula. Under the Innovative Assessment Demonstration Authority (IADA), Louisiana partnered with NWEA to pilot such an approach in grades 4–7, explicitly integrating state assessments with two specific curricula: **Wit & Wisdom** and **Guidebooks**.

In Louisiana's model, test items were drawn from the actual texts and topics in the curriculum modules—novels, informational texts, and coherent text sets that students encounter during instruction. All test items asked students to analyze characters, trace themes, interpret arguments, and write evidence-based responses **about those texts**, not about random passages detached from their classroom learning.

Early feedback was strikingly positive. Students, teachers, and parents **preferred** these curriculum-based assessments to the standard, curriculum-agnostic tests, and students completed the tests at a higher rate than they did the traditional state tests. Students reported that these tests were more authentic, more meaningful, and fairer. Teachers reported that results were easier to interpret and connect back to instruction, because they aligned directly with what had been taught.

These benefits were not accidental; they flowed from basic principles of assessment validity and equity:

- **Validity.** If the intended construct is "students' ability to comprehend and analyze grade-level complex texts as taught through a curriculum," then the most valid test is one **that samples from that curriculum's text set**. This is entirely consistent with the *Standards for Educational and Psychological Testing* and widely accepted alignment frameworks.
- **Equity**. Curriculum-based assessments reward **engagement with the taught material**, which schools and districts can influence, rather than differential exposure to cultural capital outside of school, which they cannot. Curriculum-aligned tests are better positioned to reduce, rather than amplify, the relationship between test performance and socioeconomic status.
- **Instructional usefulness.** When tests measure understanding of the actual taught texts and topics, their results become **immediately actionable**. Teachers can see which units and concepts students grasped and which need reteaching; they do not have to infer from decontextualized skills scores. This is particularly true in a "rolling" for-stakes assessment in which there are three test windows, each of which provides instructional feedback.

Louisiana's IADA work demonstrates that curriculum-integrated assessment is both feasible and promising. But it also surfaced the central barrier to scaling this model: the **cost and complexity** of building and maintaining high-quality item banks for multiple curricula across multiple grades, with development costs of more than $1million per grade level for a single curriculum.

A second barrier has been the policy environment. In the past, states have assumed that USDOE would not allow a state to satisfy federal law and regulations through multiple tests (this despite existing situations such as those in New York's Portfolio Districts). But with recent changes and calls for innovative assessments, it may now be possible for a state to operate **more than one ELA assessment**, as long as each meets stringent technical criteria and is comparable for accountability purposes.

One additional factor makes this approach more plausible now than ever before, which is the success of the HQIM movement and the Knowledge Matters campaign in persuading states and systems to adopt stronger ELA materials.

Policy Opportunity: Multiple Curriculum-Based Assessments in ELA

The factors above conspire to create a powerful opportunity:

- A state could approve a set of **high-quality, knowledge-rich ELA curricula** (e.g., Wit & Wisdom, Guidebooks, EL Education, Core Knowledge Language Arts);
- For each approved curriculum, the state, or a consortium of states, could approve the use of **curriculum-specific ELA** assessments designed by a third party, tightly aligned to that curriculum's texts and learning progressions.
- District, charter, private, and in some states, home schools choose among approved curricula as they do today, and their students would take the assessment matched to the curriculum taught in their school districts.

In effect, we would move from a single, curriculum-blind test to a **portfolio of curriculum-based assessments that shared the same design**, each psychometrically robust and comparable at the scale-score level, but anchored in distinct, coherent bodies of content.

This vision is well-aligned with IADA's intent to support **instructionally embedded, innovative assessments** that better capture deeper learning. The key question is whether such a system can be built and sustained in a cost-effective, technically defensible way.

AI as Enabler: Human–AI Co-Design of Curriculum-Based Item Banks

Historically, building high-quality items at scale is slow and expensive. For each curriculum, developers must generate multiple-choice and open-response items across grades and modules, review them for alignment and bias, pilot them, analyze their psychometric properties, and continuously refresh the bank to maintain security. The LDOE model also required the scoring of essays and content knowledge – AI may now be able to handle this.

Recent advances in **artificial intelligence**, particularly large language models (LLMs), materially change this calculus—but only if AI is used in a disciplined, expert-governed way. The pragmatic and plausible model is **human–AI co-design**.

Design Principles:

1. **Specify the content domain with precision**. For each grade and curriculum module, the designers establish the allowable item formats, plausibly including the three used in the Louisiana IADA assessment: comprehension of the anchor text of the unit, a topic-related "warm read" that the student would be asked to compare to the anchor text, and – depending on the grade level - an assessment of the knowledge acquired from the unit as a whole through writing tasks.
2. **Use AI to generate draft items at scale**. Given the texts and item specifications, AI systems can produce large numbers of draft multiple-choice items (with keys and plausible distractors) and open-response prompts (with draft rubrics and exemplar responses) that directly reference the curriculum texts.
3. **Apply rigorous human review and editing**. ELA content experts, bias reviewers, and psychometricians then review, refine, and select from these drafts. AI's role is to accelerate ideation and drafting; humans ensure validity, clarity, accessibility, and fairness.
4. **Pilot and calibrate.** Selected items are field-tested with students who have studied the relevant curriculum. Their performance is analyzed using classical and modern psychometric methods to calibrate difficulty and discrimination and to detect differential item functioning.
5. **Construct secure, calibrated item banks.** Only items that pass both expert review and empirical testing enter the operational pool. These banks are then available for use in summative tests and aligned interim assessments.

The process outlined above is not speculative. Earlier generations of automated item generation have shown that, with well-designed templates and constraints, it is possible to create psychometrically sound items efficiently. Major assessment organizations are already piloting LLM-based item drafting and automated scoring in low-stakes contexts. What is new is the possibility of applying these techniques **systematically to curriculum-specific domains**, thereby making multi-curriculum ELA assessment technically and financially feasible.

Proposed Next Steps

To seize this opportunity, we propose that the U.S. Department of Education and interested states take the following steps:

1. **Clarify and communicate federal flexibility**. USDOE can reaffirm, through guidance and technical assistance, that states may operate **multiple ELA assessments** aligned to different state-approved curricula, provided they meet specified peer-review standards for technical quality and comparability.
2. **Support a multi-curriculum pilot.** Support from publishers, foundations, and states will enable us to:

    - Extend the Louisiana model to additional grades and content-rich curricula;
    - Build and validate AI-assisted, curriculum-specific item banks following rigorous human–AI co-design protocols;
    - Conduct independent studies of **validity, equity, and instructional impact** comparing curriculum-based and traditional ELA assessments; and

- Training for state and system leaders in interpreting and using results from curriculum-based assessments.

3. **Engage practitioners.** Classroom teachers and school leaders should be engaged in design and feedback cycles to ensure that assessments remain instructionally useful and feasible to implement.

Conclusion

The United States has invested heavily in the adoption of high-quality, knowledge-building ELA curricula. Unfortunately, our state assessments still reflect an outdated, skills-only view of reading that is misaligned with both research and practice. Curriculum-based ELA assessments—anchored in the texts students actually study—offer a more valid, more equitable, and more instructionally powerful alternative.

What has long seemed impractical at scale is now within reach. The combination of **federal flexibility**, **state leadership**, **publisher support**, and **carefully governed AI-assisted item development** can enable a new generation of ELA assessments that truly measure what matters, support better teaching, and give all students a fairer chance to demonstrate what they know and can do.

Louisiana's experience shows that students, educators, and families are ready for this change. The next step is to build the policy and technical infrastructure to take this model from promising pilot to national practice.

Appendix A:

A short overview of work on LLM Assessment models (certainly not fully up to date since the field is changing quickly).

1. ETS (TOEFL, GRE, etc.)

Automated scoring

- ETS has used automated scoring for years (e-rater, c-rater). Recently, they've begun integrating more advanced NLP / LLM-style models into scoring pipelines.
- ETS research papers since ~2022 discuss transformer-based models for:
- Essay scoring
- Short-answer scoring
- Speaking assessment (e.g., automated speech scoring in TOEFL practice products)

Item generation

- ETS has been exploring AI-assisted item authoring, including:
- Using large language models to draft reading passages and multiple-choice items
- Human review/editor workflows where AI suggests stems, distractors, or variations on existing items

(For high-stakes exams, items are still heavily reviewed and often only AI-assisted rather than fully AI-authored.)

2. Pearson

Automated scoring

- Pearson's **Intelligent Essay Assessor (IEA) and Versant** (for speaking) have evolved toward deep learning and transformer-based architectures.
- Recent Pearson technical communications describe:
- Neural-network–based writing and speaking scoring
- Experiments with foundation models/LLMs to improve feedback and scoring robustness

Item drafting

- Pearson has presented **on AI-assisted item generation** at assessment conferences (NCME, ATP, etc.):
- Using generative models to create draft items aligned to existing blueprints
- Using AI to generate distractors, text passages, and variant forms of items for pretesting

3. Curriculum Associates (i-Ready)

- Public talks and blog posts from Curriculum Associates describe:
- Using large language models to **draft formative assessment items aligned** to standards
- Restricting use to low- or medium-stakes settings at first, with psychometric review
- Their R&D has focused on:
- Controlled item templates fed to LLMs

- Human-in-the-loop editing and bias checking

4. ACT

Item authoring pilots

- ACT researchers have reported pilot work where LLMs are used to:
- Generate reading passages and stems matching ACT content specifications
- Create "clones" or variants of existing items for experimental use
- These are typically **pilot or research-only** and go through editorial and psychometric review before any operational use.

Scoring research

- ACT has explored transformers for:
- Automated essay scoring
- Short constructed-response scoring in STEM areas

5. College Board

- For the SAT and AP programs, College Board has been cautious but is:
- Investigating LLMs for **item drafting support (**e.g., helping content developers brainstorm prompts, distractors).
- Exploring AI-based scoring of **practice essays and short responses** in online prep environments.
- Given the stakes, they tend to keep LLMs in supportive / pre-production roles with extensive human vetting.

6. State & District Assessment Vendors

Several K–12 assessment vendors and consortia have talked publicly about pilots:

- **Smarter Balanced / SBAC ecosystem vendors**
- Experiments using LLMs to draft ELA items and reading passages.
- Work on automated scoring of short-constructed responses using transformers.
- **NWEA (MAP)**
- Blog posts and webinars describe R&D using LLMs to:
- Generate item prototypes aligned to RIT scale regions
- Provide automated writing feedback and scoring in practice environments
- **Smaller formative assessment platforms** (e.g., EdTech startups)
- Many now market features like "AI-generated quiz items," "AI writing feedback," or "instant AI scoring" based on LLM APIs (OpenAI, Anthropic, etc.). These are typically:
- Formative/low-stakes
- Teacher-controlled (teachers accept/edit items and scores)

7. Higher Ed / MOOC Platforms

- **Coursera, edX, and similar platforms**
- Use LLMs to draft quiz questions and practice problems from course materials.

- Pilot **automated short-answer and coding-answer scoring**, with instructors able to override.
- **Institutional LMS integrations (Canvas, Blackboard, etc.)**
- Add-ons now allow instructors to:
- Auto-generate quiz items from lecture notes/slide decks
- Use LLMs to pre-score open-ended responses or flag those needing human review

8. Research & Prototypes (not yet fully operational)

Even where there's no large-scale deployment, there's a lot of published work:

- **Automated Item Generation (AIG) with LLMs**
- Using GPT-style models to generate math word problems, vocabulary items, reading comprehension questions.
- Template-based prompts (e.g., "Generate a 4-option multiple-choice item about X, aligned to standard Y, with difficulty Z").
- **Automated Scoring with Transformers/LLMs**
- Many NCME/AEA/ATP papers on replacing or augmenting traditional AES engines with BERT, RoBERTa, and newer LLMs.
- Focus on:
    - Domain adaptation (fine-tuning by prompt or LoRA)
    - Fairness/bias auditing
    - Score stability and explainability.

Technical Appendix B: Test Design AI-Assisted Item Generation for Curriculum-Based ELA Assessments

Overview and Design Goals

Objective: Develop and maintain high-quality, curriculum-aligned ELA assessments (grades 4–7) using a **human–AI co-design pipeline** that:

- Ensures **validity** and **alignment** to specific curricula (e.g., Wit & Wisdom, Guidebooks).
- Supports **equity** by assessing the actual texts and topics that students study.
- Uses AI to increase **efficiency** and scale of item development without compromising quality.

Content Domain Specification

For each **grade–curriculum–module** combination (e.g., Grade 5, Wit & Wisdom, Module 2):

Define the domain of texts

- Anchor texts, supporting texts, and key excerpts used for instruction.
- Thematic and knowledge domains (e.g., American Revolution, ecosystems, civil rights).

Map standards and learning targets

- Link each text/module to specific standards (e.g., RL.5.1, RI.4.3, W.6.1).
- Clarify the cognitive targets: inference, analysis of structure, author's craft, theme, argument, etc.

Develop item specifications (item specs)

- For each planned item type:
- Target standard(s)
- Depth of Knowledge (DOK) level
- Response format (MC, multi-select, evidence-based selected response, short constructed response, extended response)
- Constraints (readability, word count, use of quotations from the text, etc.)

These specs define the **construct** that AI-generated items must operationalize.

AI-Assisted Item Drafting

Inputs to the AI system

- Text or text excerpt (with citation and location in curriculum).
- Grade level and student population.
- Target standard(s) and DOK level.
- Item type and stem/distractor/rubric requirements.
- Style and accessibility constraints (plain language guidelines, avoidance of bias triggers, etc.).

Outputs from the AI system

Multiple-choice item drafts

- Stems that directly reference the text and target the specified standard.
- One keyed correct answer, with justification linked to textual evidence.
- Three (or more) distractors that are:
- Plausible but incorrect,
- Derived from common misconceptions or partial understandings,
- All similar in length and tone to avoid cueing.

Short- and extended-response prompts

- Prompts requiring evidence-based reasoning about the text(s).
- Clear descriptions of what an excellent response should demonstrate (for rubric alignment).

Proposed scoring rubrics

- Analytic rubric dimensions (e.g., comprehension/analysis, use of evidence, organization, language).
- Performance level descriptors (e.g., 0–4 scale) with sample phrases and exemplar responses.

Item variants

- Multiple versions of each item with controlled variation in:
    - Wording,
    - Difficulty,
    - Focus (e.g., theme vs. character motivation) within the same text.

Human Expert Review and Refinement

Step 1: Content and construct review

- ELA content experts and experienced item writers:
- Verify alignment to standards and curriculum.
- Check that the item truly requires the intended cognitive process (not superficial recall).
- Ensure correct answer is unambiguously best; distractors are plausible but clearly wrong.

Step 2: Fairness, bias, and accessibility review

- Specialists in bias and accessibility:
- Screen for cultural or socioeconomic bias beyond the inherent content of the curriculum text.
- Check language complexity and clarity against grade-level readability and accessibility guidelines (including accommodations for English learners and students with disabilities).

Step 3: Editorial and technical refinement

- Editors and psychometricians:
- Harmonize style and formatting across items.
- Confirm adherence to item specs (length, vocabulary constraints, etc.).
- Flag items that may be too easy or too difficult based on expert judgment and text demands.

Field Testing and Calibration

Pilot Testing

- Administer draft items to representative samples of students **who have studied the relevant curriculum module.**
- Collect:
    - Item response data (for MC and selected response).
    - Samples of student written responses (for constructed-response items).

Psychometric Analysis

- Use classical test theory (CTT) and/or item response theory (IRT) to estimate:
- Item difficulty and discrimination.
- Differential item functioning (DIF) across subgroups (e.g., race/ethnicity, gender, socioeconomic status, English learner status).
- Flag items that:
    - Show poor discrimination.
    - Have unexpected DIF not attributable to curriculum exposure.
    - Appear misaligned with intended difficulty levels.

Cognitive Labs

- Conduct small-scale think-aloud sessions with students familiar with the texts.
- Verify that students interpret stems as intended and that the item elicits the targeted reasoning.

Item Bank Construction

Only items that pass content, fairness, and psychometric reviews are admitted to the operational **item bank**, tagged with:

Curriculum, grade, module, and text.

Target standard(s) and DOK.

Calibrated difficulty parameters.

Automated and Human Scoring of Constructed Responses

Scoring Model Development

- Collect a **training** set of student responses scored by expert human raters using the rubric.
- Train AI scoring models to predict rubric scores.
- Evaluate:
    - Exact agreement and adjacent agreement with human raters.
    - Differential performance across subgroups.

Operational Use

- For accountability use, AI scoring is:
- Either supplemental (e.g., pre-screener or second reader);
- Or combined with human scoring (e.g., AI + human adjudication for contentious scores), depending on state policy appetite.

Governance, Security, and Continuous Improvement

- **Governance**: Establish an assessment technical advisory committee including psychometricians, ELA scholars, curriculum developers, and equity experts.
- **Security**:
  - Treat AI-generated items as secure content once accepted into the bank.
  - Implement strict item exposure controls and rotation policies.
- **Continuous Improvement**:
  - Periodically re-run item statistics to check for drift.
  - Use AI to propose revisions or replacements for underperforming items.

This appendix positions AI clearly as a **tool under expert and psychometric control**, not as an autonomous test designer.

Appendix C: Selected References

A. Knowledge, reading comprehension, and curriculum

Foundational theory / cognitive science

- **Hirsch, E. D. (2003).** *Reading comprehension requires knowledge—of words and the world.* American Educator. – Central argument: background knowledge is critical to reading comprehension; "skills-only" approaches misdiagnose the problem.– Useful because it directly supports your critique of text-agnostic, "skills-based" tests.
- **Hirsch, E. D. (2016).** *Why Knowledge Matters: Rescuing Our Children from Failed Educational Theories.* Harvard Education Press. – Book-length argument for knowledge-rich curricula; includes discussion of assessment implications.
- **Recht, D. R., & Leslie, L. (1988**). *Effect of prior knowledge on good and poor readers' memory of text.* Journal of Educational Psychology, 80(1), 16–20. – Famous "baseball study": poor readers with high domain knowledge often outperform strong decoders with low knowledge on comprehension tasks. – Empirical anchor for claim that text-agnostic tests privilege students with more background knowledge.
- **Cromley, J. G. (2005).** *The role of vocabulary and background knowledge in reading comprehension.* (See Cromley & Azevedo 2007, Contemporary Educational Psychology, 32(3), 349–382.)– Meta-analytic / structural modeling work showing substantial contributions of background knowledge to reading comprehension.
- **Willingham, D. T. (200**6). *How knowledge helps.* American Educator. – Teacher-friendly synthesis of cognitive science on knowledge and comprehension; heavily cited in curriculum reform.

B. Curriculum and outcomes

Even when not explicitly about assessment, these show that **knowledge-rich, text-based curricula** outperform generic skills programs.

- **Steiner, D. (Ed.). (2017**). *Curriculum Research: What We Know and Where We Need to Go.* Johns Hopkins Institute for Education Policy. – Synthesizes evidence that high-quality curriculum adoption is associated with improved outcomes.
- **Polikoff, M., & Dean, J. (2019).** *The Supplemental Curriculum Bazaar: Is What's Online Any Good?* (and related work on curriculum alignment). – Shows large variation in quality and alignment; implicit case that assessments must be coherently tied to strong curricula.
- **Chingos, M., & Whitehurst, G. (2012).** *Choosing Blindly: Instructional Materials, Teacher Effectiveness, and the Common Core.* Brookings. – Argues that curriculum choice matters and is underleveraged; helpful for framing why state-level curriculum-aligned assessment is a missed opportunity.

Evidence on content-rich ELA curricula

- **Cabell, S. Q., Petscher, Y., Dyer, S., & Justice, L. M. (2019).** Sh*ared book reading interventions with preschool children: A meta-analysis*. Review of Educational Research. – Shows that content-rich, text-centered instruction improves language outcomes; supports the logic that tests should sample those texts.
- **McEachin, A., & Atteberry, A. (2017).** Studies of EL Education / Expeditionary Learning ELA curriculum (various reports). – Document positive effects of adopting a coherent, knowledge-based ELA curriculum.
- **Research on Core Knowledge Language Arts (CKLA)** by the **University of Virginia** and others. – Shows stronger reading outcomes where CKLA is implemented; you can use this to argue that content-based tests would more fairly capture what students are learning.

(For a more state-facing synthesis, the **Knowledge Matters Campaign** often cites and aggregates these.)

B. Assessment validity, alignment, and curriculum-embedded assessment

Core validity and alignment frameworks

- **AERA, APA, & NCME. (2014***). Standards for Educational and Psychological Testing*. – Gold standard on test validity. – You can cite the core principle: validity depends on alignment between the test and the intended construct (here, comprehension of grade-level complex texts as taught).
- **Webb, N. L. (1997, 2007).** *Criteria for alignment of expectations and assessments.* – Introduces alignment methodology widely used by states; supports your argument that tests should align not just to abstract standards but to the actual enacted curriculum.
- **Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001**). *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academies Press. – Strong conceptual grounding for aligning assessment with cognition and instruction (the "assessment triangle").

Curriculum-embedded and instructionally useful assessment

- **Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013).** *Criteria for High-Quality Assessment.* Stanford Center for Opportunity Policy in Education. – Advocates assessments that are instructionally embedded and closely connected to curriculum and teaching.

- **Marion, S., & Shepard, L. (2021).** Various Center for Assessment briefs on innovative assessment systems. – Discuss how assessment systems can integrate instruction and testing, including curriculum-embedded components.

C. Equity, background knowledge, and test bias

- **Ferguson, R. (1998, 2008)** and others on achievement gaps and standardized tests (various reports). – Provide evidence that generic standardized tests often correlate strongly with socioeconomic status, in part via background knowledge pathways.
- **National Research Council (2012).** *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century.* – Emphasizes that deeper learning is context- and content-dependent; supports argument that tests must be grounded in substantive content domains.

D. AI, item generation, and automated scoring

- **Gierl, M., Lai, H., & Turner, S. (2012**). *Using automatic item generation to create multiple-choice test items.* Medical Education, 46(8), 757–765.– Classic work demonstrating that automatic item generation can produce psychometrically sound items when templates and constraints are well defined.
- **Arendasy, M., & Sommer, M. (2012).** *Item generation and automatic item generation in psychometrics. Psychological Test and Assessment Modeling.– Technical discussion of parameterized* item generation and its benefits.
- **von Davier, A. A. (Ed.). (2019).** Handbook of Automated Scoring: Theory into Practice. CRC Press. – Focused more on scoring, but relevant to AI in assessment systems more broadly.

LLM-based question generation & scoring (more recent)

- **Research and white papers from ETS, ACT, NWEA, and Pearson** on AI-assisted item development and automated scoring. Many have public briefs documenting pilot projects where AI:
  - Generates draft items; and/or
  - Scores short constructed responses.