NCSA 2026

WestEd

Responsible AI in Education

Another update

Sarah Quesen, PhD, MPH

Senior Director of Assessment

WestEd

At WestEd, we help education leaders make sense of **measurement and assessment,** from **rapidly emerging technology to shifting state and federal policy.**

We are committed to *measuring what matters.*

- Tools that support learning
- Evidence that drives good decisions
- Fairness throughout the system

WestEd.org/Assessment

What are we doing with AI?

- Content generation from email to assessments
- Feeding transcripts, long PDFs, or chaotic meeting notes into a model to summarize
- Using AI as a sounding board to generate ideas or troubleshoot problems
- Write code, debug programs, generate technical documentation

Can AI help us build assessment content that reflects all learners?

- Assessment content development is slow and expensive
- The default has always been to strip passages of context, aim for "neutral“

We wanted to see if AI could produce content that uses context rather than avoiding it to engage and reflect students.

WestEd

2023-2024: Early Efforts

- Built a RAG pipeline - fed it passage writing guidelines, exemplary passages, Common Core standards, CRA guidelines, and a few shot example
- No difference: custom: 70% vs null: 73% average rubric score

2024-2025: Fine Tuning

- Retrained the model with curated examples in a secure environment tagged with Lexile range, reading standards, depth of knowledge
- What improved: Grade level specs, word counts, format matched developer needs. What did not improve: Generated content's responsiveness

In Progress: Contextual Transformation

Instead of generating responsive content from scratch, adapt existing vetted content to different local contexts.

- Take a passage that works. Change the setting. Shift the cultural backdrop. Keep the measurement intent.
- Potentially help scale shared item banks across state contexts
- Planned research to investigate if item properties are transferable

Other AI projects

- Building machine-learning models to solve complex scale comparability problems
- Training a content alignment engine
- Generating sample responses based on item characteristics
- Investigating structural bias in LLM output
- Modernizing approaches to AI scoring evaluation for both speech and text
- And more! I even built an ask-the-syllabus chatbot for my statistics class this summer.

WestEd

As researchers, we report our limitations and describe our tools and measures honestly.

AI should be no exception.

Hidden Costs

Environmental

- In 2025, utilities received data center connection requests totaling about 700 gigawatts, more than all the electricity the entire US used in 2023
- Texas data centers alone used about 49 billion gallons of water in 2025, projected up to 399 billion by 2030.
- Meta's Hyperion and Microsoft's Fairwater are each headed toward about five million high end GPU equivalents when they reach full capacity in late 2027 to early 2028.

Human

- Kenyan workers paid $2/hour to tag and label deeply disturbing content so AI systems can learn what to filter out.

What is AI doing with us?

- Confident, fabricated facts and citations
- Agreement (the customer is always right)
- Bias carried from the training data
- Homogenized, averaged output
- Deepfakes and synthetic media

Messick defined a unified validity as a judgment about how scores are interpreted and used, with consequences as part of that judgment rather than an optional audit.

Consider: validity as a judgment about how AI tools are interpreted and used, with consequences as part of that judgment rather than an optional audit.

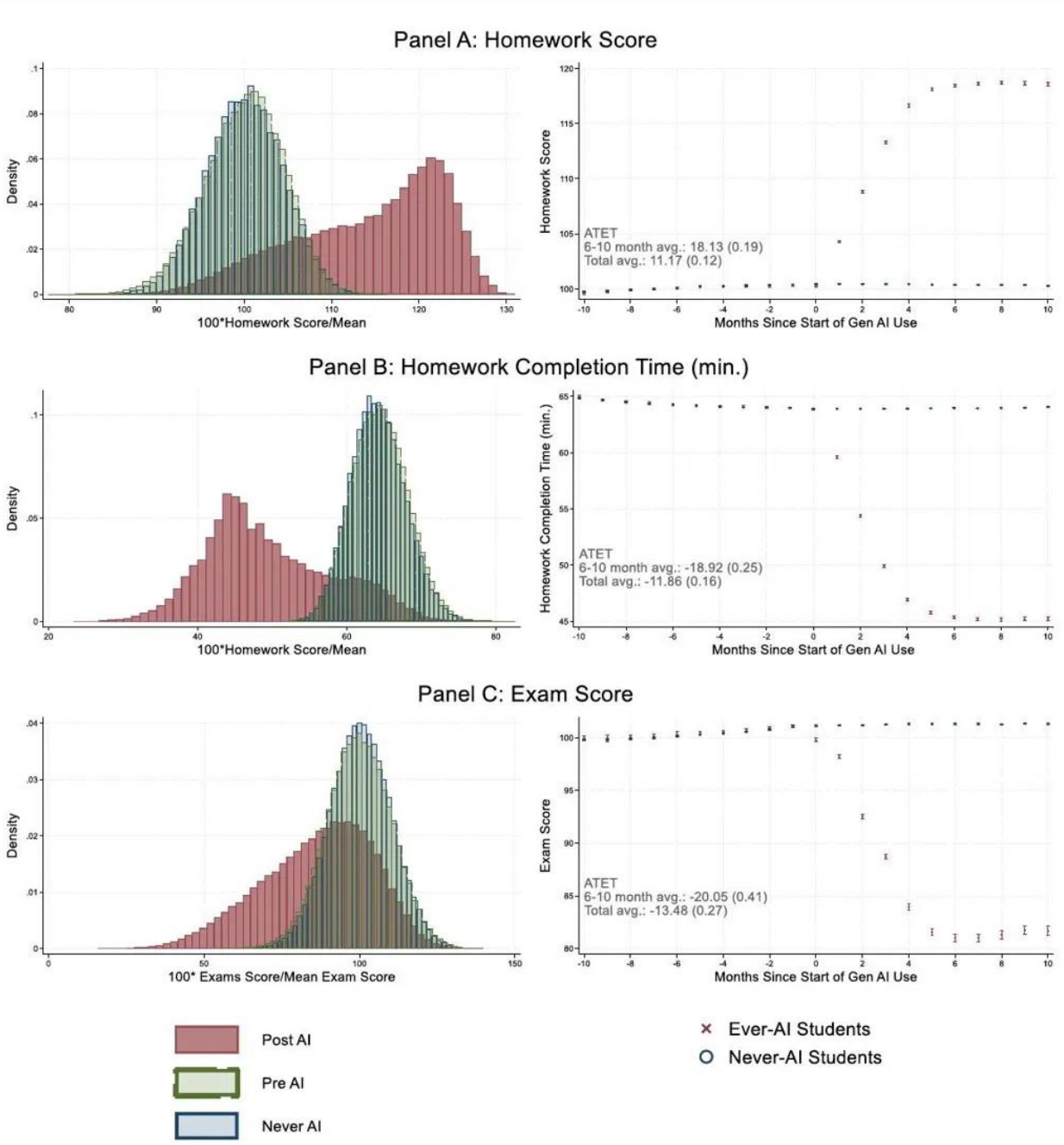
Evaluating social consequences is a fundamental requirement of establishing validity.

What are the consequences?

- **Performance without mastery**
  - Generative AI lets people raise their performance without mastering the underlying knowledge and skills.
- **Learning needs the struggle AI removes**
  - When an AI tool resolves the difficulty before the student notices it, the metacognition, the self-efficacy, and the durable skill never form.

WestEd

- Tools have been shown to raise performance while students have access, then the gains fade once the tool is gone.
    - That fade is a failure to transfer durable skill (Stanford SCALE review, Fesler et al. 2026).
- With AI, homework scores rise and homework time drops, but exam scores fall (Stromberg, Lei & Wu 2026).



When the product is free, you are the product.

- Commercial products that prioritize engagement and are designed to keep users chatting
- Our brains favor the low effort path and prefer tools that make learning feel easier
- AI is built to maximize engagement, to agree, and to resolve the difficulty for us
- The struggle that builds learning is the first thing both sides remove

So, what can we do?

So, what can we do?

Ask a lot of questions.

- When someone promises an AI can generate amazing content with no human oversight
- You hear claims about student learning with AI that seem too good to be true
- When you find yourself pivoting to AI to write your LinkedIn post or email
- When sensitive or personal data might go into a tool you do not control
- When the human relationship is the point

Learn More at WestEd.org