What AI Can—and Cannot—Tell Us About What Students Know

In a conversation with Dr. Brent Duckor and Dr. Carrie Holmberg, assessment researcher Dr. Sarah Quesen explains why the arrival of generative AI makes old questions about purpose, validity, bias, feedback, and student thinking more urgent—not less.

Artificial intelligence has entered education with a compelling promise: assessments can be created more quickly, scored more efficiently, and translated into immediate, personalized feedback. Yet speed and fluency can disguise a fundamental problem. Before asking what an AI system can produce, educators must ask what an assessment is for, what it measures, how results will be used, and what consequences may follow.

That was the central message of Dr. Sarah Quesen, senior director of assessment at WestEd, during an hour-long conversation with Dr. Brent Duckor and Dr. Carrie Holmberg for the Center for Innovation in Applied Education Policy's Thinking Through AI series. Drawing on her work in psychometrics, automated scoring, machine learning, and assessment development, Quesen described a field moving rapidly—but not always thoughtfully—toward AI-mediated testing and feedback. That distinction became the organizing thread for the discussion and its implications for schools.

Quesen is not anti-technology; she is actively leading research on AI in assessment. But impressive output should not be confused with valid measurement. AI can generate text that resembles an assessment passage, score report, or teacher feedback. Whether it supports defensible conclusions about what a student knows is another question.

When plausible output obscures purpose

Psychometricians have long begun with questions of purpose, interpretation, and consequence. Quesen worries that AI is sometimes treated as a special innovation that deserves a pass from those expectations. Schools and systems are being presented with polished products, personalized dashboards, and confident claims before anyone has adequately established the construct being measured or the actions educators should take from the results. Dr. Quesen notes:

> *"Educational decisions are being made on Star Trek assumptions with Flintstones tools."*

The metaphor captures the gap between public expectations and current capability. Large language models do not reason independently in the way many users imagine. They predict likely sequences of words from enormous bodies of human-produced text. Because those data reflect unequal access, historical stereotypes, and the uneven record of the internet, the systems can reproduce dominant patterns while making their output sound authoritative.

Quesen described this predictive tendency as a movement toward the middle: a regression to the most likely, conventional, or mediocre response. In assessment, that matters. The question "What does this measure?" can quietly be replaced by "What does the model predict?" Those are not equivalent questions.

Duckor connected this concern to basic assessment design. A measurement model belongs inside a framework that defines the construct, guides item design, specifies scoring, and supports interpretations. Without that architecture, AI can produce reports without producing meaningful measures. Quesen called the result "death by a thousand dashboards." Her practical test is simple: What would a teacher do with this?

What happened when AI tried to write assessment passages

Quesen's team at WestEd has spent roughly three years investigating whether AI can support the development of English language arts passages. The initial expectation was optimistic. A model supplied with standards, item-writing specifications, and a culturally responsive assessment framework might generate richer, more varied passages.

The results were sobering. A retrieval-augmented model supplied with carefully curated materials did not produce passages that human reviewers judged meaningfully better than those generated without the additional resources. Fine-tuning improved compliance with clearer technical requirements, including format, length, grade-level features, and standards alignment. It did not reliably produce culturally responsive content or the nuanced representation of learners in varied contexts that the researchers sought. The passages often became sentimental, generic, or artificial.

Quesen has not abandoned the work. Her team is exploring whether AI can vary the contexts of complex science tasks without disrupting their measurement properties. That use may be promising because the constraints are more explicit. Still, the broader lesson and quest for sound judgement (APA, AERA, NCME, 2014) based on AI infused test data remains:

> *"The hard parts—the parts that require human judgment—did not get easier."*

AI performed best on features that could already be specified as rules. It struggled with features that depend on interpretation: what a passage communicates to a particular learner, how a context may be experienced, and whether content represents students respectfully and meaningfully. At

the same time, the existence of a fast and inexpensive tool creates pressure to reduce the human review that remains essential. When a passage looks acceptable at first glance, it becomes harder to defend the time and cost required for expert judgment, educator review, field testing, and revision.

Bias can enter before, during, and after scoring

Quesen's research has also examined whether automated scoring systems evaluate students comparably across race, gender, language background, and disability status. Bias can enter at many points. It may be embedded in the task or rubric, reflected in the papers selected to train a scoring model, or introduced through the examples and expectations used to prepare human raters.

Earlier scoring systems raised the concern that machines might reproduce human bias. Generative AI adds another risk: models can introduce stereotypes from broader training data even when student work is unchanged. Quesen discussed research in which identical eighth-grade essays received different feedback when demographic labels were altered. The models' responses reflected patterned assumptions about grammar, responsibility, emotionality, and directness.

That is especially concerning because AI feedback is fluent and confident. A biased human rater can be identified through comparisons, disaggregated results, retraining, and monitoring against established standards. A generative model may function as a black box, sending polished feedback directly to a learner before anyone examines whether different students are being held to different expectations.

Quesen drew a useful distinction between high-stakes summative assessment and the rapidly expanding formative technology market. Established summative programs generally retain multiple layers of content review, field testing, statistical analysis, and checks for differential item functioning. In some classroom products, however, content or feedback can travel directly from the model to the student. For Quesen, that less-regulated pipeline may pose the more immediate danger.

Feedback is not a logistics problem

Duckor connected those concerns to interviews they conducted with teachers, many of whom described AI-generated responses as a "feedback firehose." More comments did not necessarily mean more useful guidance. The output often failed to account for a learner's zone of proximal development, the instructional moment, or the particular revision a student was ready to make.

Quesen agreed that large language models tend toward generic rather than genuinely individualized responses. They also are designed to sustain engagement and often validate users rather than challenge them. That tendency can work against the productive tension good feedback sometimes requires.

As Duckor and Holmberg (2023) note: For teachers, feedback is not merely text attached to work. It carries knowledge of the learner, disciplinary expertise, expectations, trust, and a history of interaction. Statistically generated language may imitate that form without sharing the relationship that gives it meaning.

This is why claims of "better, faster feedback" deserve careful scrutiny. Faster may describe the delivery system, but it does not establish instructional value. Better requires evidence that the response is accurate, equitable, developmentally appropriate, and likely to help a student act.

Deeper learning depends on friction

AI may reduce logistical barriers by generating varied tasks, organizing information, or supporting carefully designed tutoring systems. But it also can undermine the construct educators hope to measure.

Deeper learning requires students to analyze, solve problems, create, persist, and revise. These capacities develop through cognitive struggle and socially meaningful work. General-purpose AI systems often remove that struggle by supplying a finished answer. Even tools that imitate a Socratic exchange may produce a generic follow-up question without truly understanding the student's reasoning, motivation, confusion, or readiness.

The measurement problem follows directly. When AI performs the critical thinking, the resulting score may indicate a student's ability to operate an AI tool rather than the student's independent capacity to reason. The construct has shifted, whether or not the assessment designer acknowledges it.

Quesen saw this in her own university statistics courses. After generative AI became widely available, grades rose abruptly and implausibly. She redesigned the course and returned some assessments to paper and controlled testing conditions—not because older methods are always preferable, but because she needed to preserve opportunities for students to do the thinking the course was intended to develop.

At the same time, Duckor noted the paradox facing educators. Students will enter workplaces where AI use is common, collaborative, and expected. Schools cannot prepare them for the future by pretending the technology does not exist. Yet students cannot use AI wisely without disciplinary knowledge, judgment, and experience with the productive struggle through which understanding is built.

Ask for evidence, not a demonstration

When a vendor presents an AI assessment solution, Quesen recommends beginning with evidence, not a demonstration or pilot anecdote. Educators should ask what the tool measures, which students were included in its development, how performance was evaluated across groups, what harms were tested, and what happens when the underlying model changes.

Even the reassuring phrase "human in the loop" requires clarification. Who is the human? What are they expected to detect? Do they have the time, training, information, and authority to identify a failure in a black-box system? Human oversight is not a safeguard merely because it appears in a diagram. Quesen noted of the challenge in AI-mediated assessment:

> *"Better according to what evidence? Faster, but at what cost?"*

AI literacy must include more than prompt-writing or confidence surveys. Students, educators, and leaders need to understand what models can and cannot do, when outputs are trustworthy, what is lost when difficulty disappears, and how to evaluate claims.

A paradigm shift without a validity holiday

Near the end of the conversation, Duckor named the larger challenge: Education is experiencing a paradigm shift (Duckor & Holmberg, 2026).

Evidence gathered before generative AI may not transfer neatly to classrooms in which machines generate feedback, mediate revision, or perform parts of the task. Even research traditions with decades of support must be revisited when the conditions of teaching and learning change.

But a paradigm shift does not erase first principles. It makes them more necessary. Purpose, construct, use, consequence, fairness, and human judgment remain the anchors of sound assessment. The task is not to choose between an AI future and an instructional past. It is to prepare students for an AI-mediated world without surrendering the relationships, disciplinary knowledge, and productive struggle that make deeper learning possible.

Duckor's words suggest the appropriate stance: recognize that the ground is moving, then resist the temptation to confuse movement with progress. In this new paradigm, AI should not receive a validity holiday. Every claim still requires evidence, every measure still requires a clear purpose, and every innovation must ultimately answer the question Quesen repeatedly brought back to the classroom: *What will a teacher and student be able to do because of it?*

Citation: Duckor, B., Holmberg, C., & ChatGPT. (2026, June 30). Thinking Through AI: AI, Educational Assessment, and the Challenge of Knowing What Students Know [Webinar summary]. Innovation in Applied Education Policy (IAEP) Center, San José State University.