SOCRATIC TESTING ENABLED BY AI

Faculty and students experiment with Northwestern-developed testing innovation designed to empower students to show what they know and professors to see it more clearly.



THE TEST IS ABOUT TO BEGIN. Students sit down at their computers. It's a familiar scene, yet the interaction about to take place between each student and the exam is revolutionary.

Rooted in ancient Greek teaching philosophy, a new Socratic testing technology enables faculty to gauge student comprehension through an AI-mediated, tailor-made conversational exam. Invented by Northwestern Engineering's Ilya Mikhelson, the software provides accurate and scalable insights into students' domain knowledge.

Like a one-on-one oral exam, the platform shapes the testing interaction in real time through a scaffolded dialogue between system and student. Mikhelson acknowledges that oral exams were his inspiration.

"Unstructured oral exams allow me to prompt, prod, and map out students' cognitive ceiling in a given topic, but the time and logistics that takes just aren't feasible, especially for a large class," says Mikhelson, Charles Deering McCormick Distinguished Professor of Instruction in the Department of Electrical and Computer Engineering. "And, while many students have interacted with computer adaptive testing, a computer-based examiner that can have an actual conversation with students was impossible until the current state of AI."

Testing the tester

In spring 2026, Mikhelson sought an opportunity for faculty and students to work together to evaluate and improve the software. David O'Neill, Charles Deering McCormick Distinguished Professor of Instruction of Biomedical Engineering, piloted the platform for a midterm exam in Fluid Mechanics (BMD_ENG 270), a required core curriculum course for all biomedical engineering undergraduates. Participating in the pilot was optional for students. Two weeks before the exam, he introduced his class of 66 students to the system's standard exam mode, presenting a live demo of its organic, personalized conversation.

Mikhelson's Socratic tester begins with low-cognitive-load questions and incrementally raises complexity with guided help along the way. The test's content and scope draw on relevant training data—lecture notes, past exams, problem sets—while instructors set guardrails through custom directions or constraints tied to their specific learning objectives. In evaluating student answers, the system is instructed explicitly to disregard spelling, sentence structure, abbreviations, or slang.



"IT IS EXCITING TO SEE FACULTY AND STUDENTS EXPLORING THE POSSIBILITIES OF NEW TESTING APPROACHES ENABLED BY AI. WHILE THE PLATFORM IS ENABLED BY AI, THE JUDGMENT, MENTORSHIP, AND HUMAN CONNECTION OUR FACULTY MEMBERS BRING TO ASSESSING LEARNING REMAIN AT THE CENTER OF HOW WE EVALUATE OUR STUDENTS."

Karen Smilowitz Associate Provost for Undergraduate Education,
James N. and Margie M. Krebs Professor of Industrial Engineering and Management Sciences

Seeking a solution, Mikhelson coded a Socratic-inspired testing platform and then partnered with the Office of the Provost to pilot the prototype in three courses in biomedical engineering and civil and environmental engineering during spring quarter 2026. The initial results, including faculty and student feedback, were promising.

Most students thought the system accurately interpreted their responses and pushed them appropriately when their answers were incomplete. Similarly, course instructors reported that the system efficiently broke down complex concepts, pushing students to articulate their reasoning rather than relying on rote memorization or factual recall.

"It is exciting to see faculty and students exploring the possibilities of new testing approaches enabled by AI," says Karen Smilowitz, associate provost for undergraduate education and James N. and Margie M. Krebs Professor of Industrial Engineering and Management Sciences. "While the platform is enabled by AI, the judgment, mentorship, and human connection our faculty members bring to assessing learning remain at the center of how we evaluate our students."

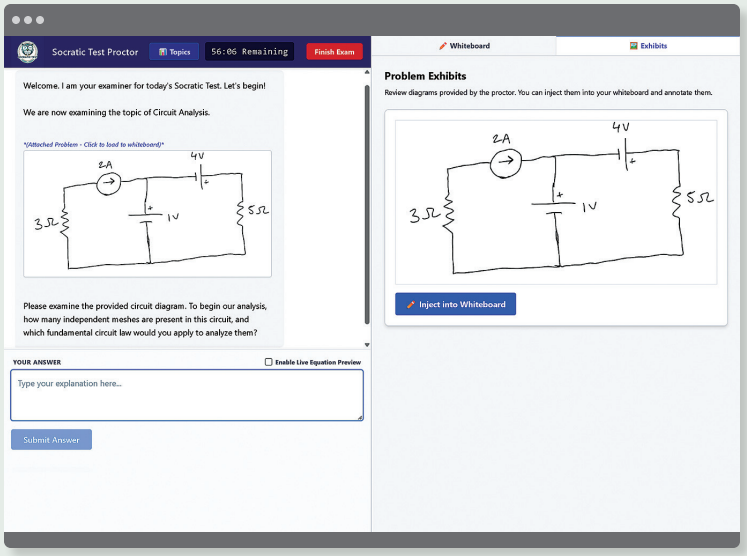
In O'Neill's course, students also had unlimited access to the platform's asynchronous practice mode, a study tool that provides detailed formative feedback on each interaction, to prepare for the midterm.

During the live exam, as a safeguard against AI hallucination, students were authorized and encouraged to challenge the tester if they identified an error or if a question posed by the system was out of bounds. The system accepted the student's assertion, flagged the interaction for audit by the instructor, and pivoted to a new question.

That is what happened with Nathan Miller, who is studying biomedical engineering and integrated science. Confident after extensive practice, he pushed back during the midterm, defending his approach to solving a problem. The system acknowledged the miscommunication and moved on.

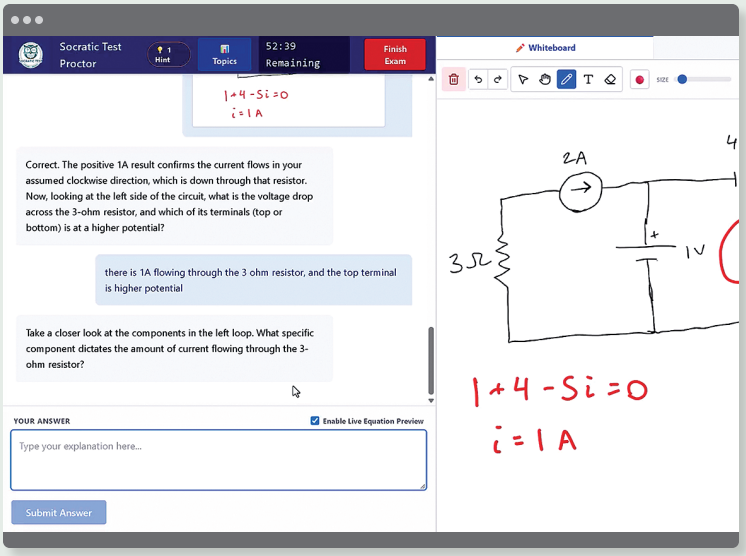
Compared to traditional exams with similar stakes, Miller reported significantly reduced stress and anxiety during the midterm using the Socratic testing platform. Overall, a majority of students surveyed after the spring pilot also reported lower stress levels.

opposite Photo by Jason Brown



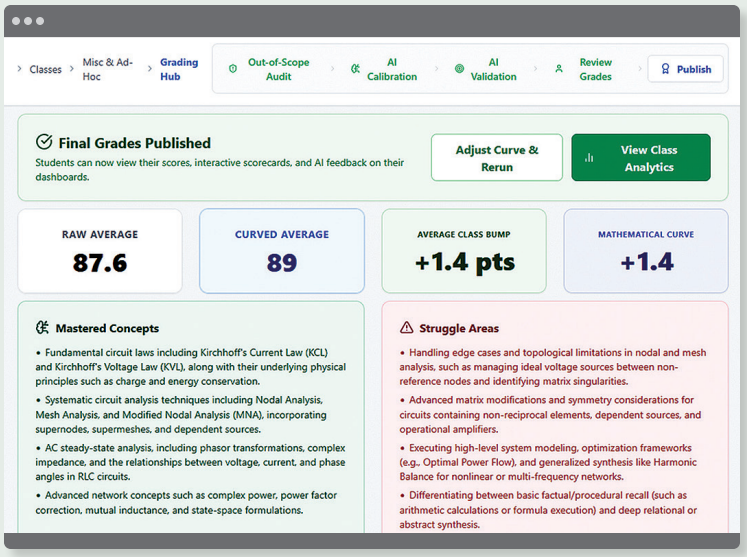
1

Because a text-based chat is often insufficient to evaluate highly complex, spatial, or mathematical skills, the tester interface includes several optional workspaces such as a whiteboard, drafting grid, integrated development environment for coding, a calculator, and reference materials like the periodic table. During the midterm exam, O'Neill's students used portable graphics tablets to write freehand, including annotating technical diagrams and formulating problems with handwritten calculus.



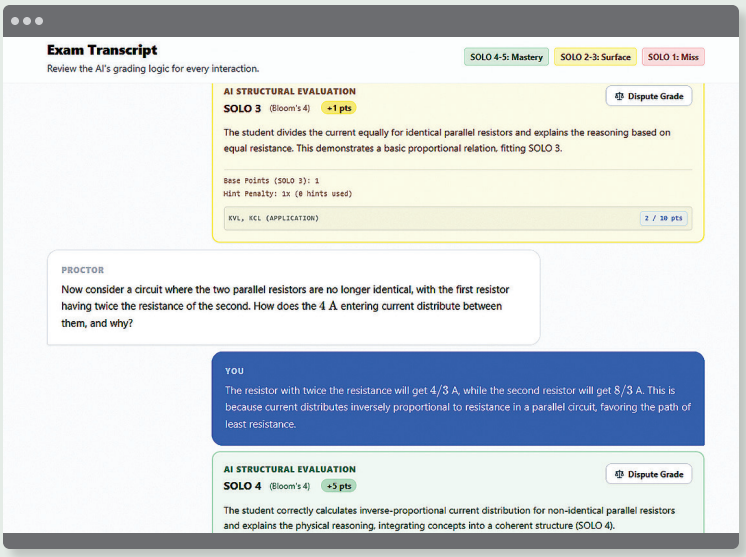
2

The testing interface shapes the testing interaction in real time through a scaffolded testing dialogue between system and student. When a student struggles, the system employs a graduated approach starting with a general nudge, as reflected in this example.



3

Based on their own scoring configuration, instructors receive detailed reports with diagnostic feedback arriving through an analytics dashboard featuring a topic-by-topic cognitive mapping of where students demonstrated subject mastery and where they struggled.



4

Once the instructor posts grades, students receive a fully marked-up transcript including point values and feedback for every response. If they disagree, they can launch an appeal directly through the platform for faculty review.

Get a tour of the Socratic testing platform.



"By recruiting classes across disciplines—from theoretical to applied, interpretive to convergent—I aim to test the universality of this system's ability to provide insights that advance deep learning."

Ilya Mikhelson Charles Deering McCormick Distinguished Professor of Instruction

"The tester really did a good job giving me layered questions, allowing me to fill in what I knew and ask for help when I fell short. I felt like I had a fail-safe if I got really stuck," Miller says. "This was such a positive experience, but it was also the first version of the software, so I'm excited to see how it can grow."

O'Neill graded each student's midterm transcript by hand and found that it helped students demonstrate integrated understanding and accurately compared their conceptual physics mastery against their mathematical fluency.

"It felt a much truer balance of what I wanted students to come away from Fluid Mechanics being able to do and understand," says O'Neill, who is also Michael Jaharis Director of Experiential Learning and associate director of the Northwestern Center for Engineering Education Research.

Mikhelson's Socratic tester was also piloted during the spring in Medical Devices, Disease & Global Health (BMD_ENG 380) and Structural Art (CIV_ENV 220). In fall quarter 2026, Mikhelson is fully integrating and evaluating the platform's human-in-the-loop grading pipeline.

Activating the test's evaluation framework

- Mikhelson designed the tester first to dictate the learning objective of each exam prompt based on the six cognitive skills of Bloom's Taxonomy—remember, understand, apply, analyze, evaluate, create. The system then applies a separate framework called the Structure of the Observed Learning Outcome (SOLO) taxonomy to measure the structural quality and complexity of the student's response.
- Instructors configure specific point capacities for both the Bloom's levels and the SOLO evaluation. During the exam, students gain points as they demonstrate evidence of knowledge.
- When an exam is over, instructors review any interactions students flagged as out of bounds. Confirmed cases are struck from every student's record for that question. For each student, Mikhelson's tester predicts a SOLO level and ambiguity score for each interaction, then sends 15 edge-case interactions to the instructor for manual review and ranking.
- Using an instructor's calibrated examples, the system analyzes a new, random set of ten interactions, and the instructor independently grades the same interactions. If the alignment is weak, the disagreements are used for recalibration, and the validation process repeats until there is strong agreement between the professor and the platform. When alignment is achieved, the system reviews the tests en masse, evaluating individual responses using the full conversational context of the exam transcript. Instructors then review the full analysis and flag any issues for recalibration and the cycle resumes.

"I want the grading process to be completely transparent for students," Mikhelson says.

- Once the instructor posts the grades, students receive a fully marked-up transcript including point values and feedback for every response. If they disagree, they can appeal directly through the platform for faculty review.
- Instructors receive detailed reports as well, with diagnostic feedback arriving through an analytics dashboard with a topic-by-topic cognitive mapping of where students demonstrated subject mastery and where they struggled.

Widening the conversation

Mikhelson is continuing to iterate and refine the tester in collaboration with an interdisciplinary research team that includes O'Neill; Jacob Kelter ('15, PhD '25), director of NetLogo and an alumnus of the Computer Science and Learning Sciences doctoral program; and Almaz Mesghina, assistant professor of instruction in psychology.

The team's plans include rigorous testing of the platform in several fall quarter 2026 courses. In addition to Mikhelson's Intro to Electrical Engineering (ELEC_ENG 202), other courses chosen for testing include Communications Systems (ELEC_ENG 307), Human Physiology (BIOL_SCI 310), Structural Steel Design (CIV_ENV 323); and First-Year Seminar: Mental Health Diagnosis and Treatment: Psychological and Economic Themes (PSYCH 101). Testing will be followed by a formal study of the software to analyze whether this testing technology helps professors accurately and equitably gauge deep understanding as compared to traditional tests.

"By recruiting classes across disciplines—from theoretical to applied, interpretive to convergent—I aim to test the universality of this system's ability to provide insights that advance deep learning," Mikhelson says.

MICHELLE MOHNEY