

Invited Editorial

If AI Can Produce the Work, What Should We Grade? AI-Durable Evidence of Competence in IS and Computing Education

Timothy C. Smith and Xihui “Paul” Zhang

Recommended Citation: Smith, T. C., & Zhang, X. (2026). Invited Editorial: If AI Can Produce the Work, What Should We Grade? AI-Durable Evidence of Competence in IS and Computing Education. *Journal of Information Systems Education*, 37(4), 448-454. <https://doi.org/10.62273/GVGA1877>

Article Link: <https://jise.org/Volume37/n4/JISE2026v37n4pp448-454.html>

Received:	September 23, 2026
First Decision:	October 2, 2026
Accepted:	October 5, 2026
Published:	December 15, 2026

Find archived papers, submission instructions, terms of use, and much more at the JISE website:
<https://jise.org>

ISSN: 2574-3872 (Online) 1055-3096 (Print)

Invited Editorial

If AI Can Produce the Work, What Should We Grade? AI-Durable Evidence of Competence in IS and Computing Education

Timothy C. Smith

Muma College of Business
University of South Florida
St. Petersburg, FL 33701, USA
smith515@usf.edu

Xihui “Paul” Zhang

Sanders College of Business and Technology
University of North Alabama
Florence, AL 35632, USA
xzhang6@una.edu

ABSTRACT

Information systems and computing educators have long inferred what students can do from what they hand in. Generative AI breaks that inference. When any student can produce a fluent schema, query, program, or report with almost no effort, the quality of a submission no longer separates those who understand from those who do not. We argue that this is a problem of assessment validity rather than academic integrity, and that prohibition, detection, and graduated-permission policies leave it unresolved. We propose three properties that make evidence of competence durable under AI assistance (substrate-tied, attention-tied, and interactively probeable) and six observational surfaces for gathering such evidence. We then consider what this asks of instructors, programs, and researchers, state what we are not claiming, and offer four testable expectations as an invitation to the community.

Keywords: Generative AI in education, Assessment, Competency, Academic integrity, Information systems education, Computing education

1. THE INFERENCE WE RELIED ON

For as long as we have taught information systems and computing, we have graded what students make. An entity-relationship diagram, a normalized schema, a set of SQL queries, a working program, a requirements specification, or a case analysis: each is an artifact, and each has served as evidence that its author can do the work. The inference from artifact to competence was never airtight, but it was serviceable, because producing a good artifact was difficult in a way that tracked understanding. A student who could not reason about functional dependencies rarely submitted a schema in third normal form.

Generative AI has made it easier to produce sophisticated work without demonstrating the knowledge and skills traditionally required to create it. Large language models produce fluent, technically sophisticated documents, code, and analyses with close to zero marginal effort, and they do so for every student, not only the capable ones. Computing educators recognized early that code generators made it far easier to produce work that passes first-year programming assignments and even exams (Becker et al., 2023), and benchmarking since then shows the capability improving quickly (Prather et al., 2023). Recent work in the *Journal of Information Systems Education* (JISE) describes the same tools helping students develop databases and understand SQL (Zhang, 2025) and write technical reports (Jiang & Nakatani, 2025). Outside the classroom, a field study of more than five thousand customer-support agents found that AI assistance helped the least experienced and lowest-skilled workers most, narrowing the measured performance gap between them and their most skilled colleagues (Brynjolfsson et al., 2025). When outputs converge in quality, inspecting the output no longer tells us who is capable.

We think this is the central assessment problem now facing our field, and we think the usual framing mislabels it. The conversation has been about academic integrity: students are using AI when they should not. The deeper problem is validity. Validity concerns the inferences we draw from a performance (Messick, 1995), and the inference from a polished submission to a student's understanding is what has failed. It fails regardless of whether any rule was broken. It fails in the courses where we permit AI precisely because our graduates will be expected to use it. So, rather than asking whether a student used AI, this editorial asks where competence remains observable when AI can produce the visible work product and what properties evidence of competence must have for us to trust it.

2. WHY THE FAMILIAR RESPONSES FALL SHORT

Four responses dominate current practice. Each is reasonable, and none tells us where competence remains observable.

Prohibition and proctored assessment restore the link between artifact and competence by excluding AI from the setting. That is defensible for foundational skills we want students to hold unaided. It cannot be the general answer in a discipline whose graduates will work alongside these tools every day.

Detection fares worse. A test of fourteen detection tools found them neither accurate nor reliable, with a bias toward classifying AI-generated text as human-written (Weber-Wulff et al., 2023). Even a perfect detector would answer the wrong question: it would tell us whether AI was used, not whether it was used competently.

Graduated-permission schemes such as the AI Assessment Scale (Perkins et al., 2024) bring welcome clarity about how much AI use a given assessment allows, and practical guidance for instructors is maturing (Gimpel et al., 2025). But a permission level is a rule about inputs. Once AI is allowed, it does not tell us what to look for.

AI literacy frameworks describe what people should understand about AI (Long & Magerko, 2020), and competency models such as IS2020 describe what our graduates should be able to do (Leidig & Salmela, 2022). Both are necessary. Neither explains how to evidence a competency when AI can produce the work products through which we have ordinarily observed it.

A fifth suggestion comes closer. It is to design assignments that require reflection and detailed documentation of the problem-solving process, as proposed in this journal and elsewhere (Zhang, 2025). We share the instinct and want to sharpen it, because a reflective essay is also an artifact, and a language model will write a convincing one on request. Asking students to show their process helps only if the evidence of process has properties that a narrative about process lacks.

3. THREE PROPERTIES OF EVIDENCE THAT SURVIVES AI

Signaling theory explains both the failure and the way out (Connelly et al., 2011; Spence, 1973). A signal distinguishes capable from less capable producers only as long as it is cheaper for the capable to produce. A polished deliverable used to be that kind of signal. Now that anyone can produce one cheaply, it carries

little information about its author. Restoring discrimination means shifting our attention to evidence whose cost still depends on the student's own judgment. We see three properties that keep that cost in place.

Evidence is *substrate-tied* when it refers to a state of the world that the evaluator can independently re-check. A student who reports testing an AI-generated query against the course database and finding that it silently drops rows with null foreign keys has made a claim the instructor can verify by running the query. The claim is either true of the database, or it is not.

Evidence is *attention-tied* when it can be produced honestly only by someone who was paying attention at the time. A decision log that marks each design choice as AI-suggested, AI-modified, or independently reached can be completed truthfully only by a student who kept track of the difference while working.

Evidence is *interactively probeable* when it holds up under follow-up questions. Why a surrogate key here? What breaks if this business rule changes? Answers remain coherent only for students who did the underlying thinking.

The argument rests on the design of the assessment rather than on limitations in current AI capabilities. These properties rely on a tie to a substrate, to an attentional state, or to a live dialog, and AI cannot substitute for any of these without leaving traces. A backfilled decision log and a fabricated verification trail fail the same checks: the claimed test was never run against the database, and questioning exposes reasoning the student never did.

Evidence with these properties takes the form of what we call process artifacts: records generated by doing the work, such as verification trails, decision logs, and framing rationales. They differ from the output artifact, whose value as a signal AI has collapsed. They also differ from after-the-fact narratives about process, which AI writes as readily as it writes the deliverable.

4. SIX SURFACES ON WHICH COMPETENCE REMAINS VISIBLE

Having identified three properties of credible evidence—being substrate-tied, attention-tied, and interactively probeable—we now consider where to look for such evidence in students' work. We propose six AI-durable competencies around which to organize that search: Problem Framing, Evidence Evaluation, Socio-Technical Thinking, Data and AI Governance, Human-AI Orchestration, and Value Realization. We use the term for observational surfaces along which competent performance can be evidenced in AI-augmented professional knowledge work. AI-durable does not mean AI-free or AI-resistant. It refers to the durability of the surface as a place to observe competence. A surface qualifies when three conditions hold: accountability for performance on it cannot be delegated to an AI system, which is not a legal person; competent performance requires contextual judgment grounded in the situation at hand; and the cost of producing credible evidence on it stays tied to the practitioner's own judgment, in the sense of the three properties above.

The first author derived the six competencies using Nickerson et al.'s (2013) taxonomy development method, iterating between the literature on generative AI, competence, and accountable knowledge work and the public record of consequential failures in AI-augmented professional practice. Table 1 presents each surface with the failure mode observed when the competency is absent and an example of durable classroom evidence.

Three features of the set matter for course design. First, the surfaces are compositional rather than independent. A single project generates evidence on several at once, and adjacent surfaces divide the terrain by stage or level: Problem Framing operates before AI is engaged and Human-AI Orchestration during and after, while Evidence Evaluation asks whether a specific output is right and Data and AI Governance concerns risk across outputs. Second, they are surfaces, not traits. The same student will perform differently under different tools, tasks, and verification cultures, so what we assess is situated performance, which is also what employers will hold our graduates accountable for. Third, the list is deliberately narrow. Communication, leadership, and negotiation matter as much as ever, but we assess them by observing the person directly, and AI has not degraded that channel. The six surfaces cover the ground where we used to rely on the artifact.

Surface	What We Observe	Failure Mode When Absent	Example of Durable Classroom Evidence (Property)
Problem Framing	How the student translates an ambiguous need into structured requirements, constraints, and criteria	Prompt Myopia: solving the specified prompt while missing the intended organizational problem	Framing rationale stating which problem formulations were considered, which were rejected, and why, defended under questioning (interactively probeable)
Evidence Evaluation	How the student interrogates AI output for accuracy, logic, and bias	Automation Bias: accepting flawed output because it appears authoritative	Verification log for AI-generated queries or code recording what was tested, against which data, and what failed; any entry can be re-run (substrate-tied)
Socio-Technical Thinking	How the student traces interactions between technical solutions and organizational systems	Isolated Optimization: technically correct solutions that fail in deployment	Impact analysis naming the specific roles, workflows, and systems in the case organization that a design changes, checkable against the case record (substrate-tied)
Data and AI Governance	How the student manages privacy, compliance, ethics, and algorithmic-bias risks	Liability Blindness: deploying solutions that violate regulatory or ethical norms	Record of the privacy, compliance, and bias checks actually run on a dataset or design, with results and any issues escalated (substrate-tied)
Human-AI Orchestration	How the student structures iterative human-AI workflows and oversight	Fragmented Workflow: ad hoc loops without strategic oversight	Decision log kept during the project, marking each choice as AI-suggested, AI-modified, or independently reached (attention-tied)
Value Realization	How the student connects AI-supported outputs to organizational objectives	Mirage of Productivity: high artifact volume without business value	Success criteria stated before the work begins, followed by an assessment of the delivered result against them (substrate-tied and attention-tied)

Table 1. Six AI-Durable Competencies as Observational Surfaces, With Failure Modes and Durable Classroom Evidence

5. WHAT OUR GRADUATES WILL BE ACCOUNTABLE FOR

The failure modes in Table 1 are not hypothetical. Eight published incidents of consequential failure in AI-augmented professional work were analyzed in developing the framework, drawn from law, customer service, hiring, consumer credit, public administration, journalism, engineering, and real estate. Each could be classified on the six surfaces without a seventh, as could four further incidents examined afterward as an out-of-sample check. The typical incident implicated two surfaces at once.

One incident makes a useful teaching case. In *Mata v. Avianca, Inc.* (2023), attorneys filed a court submission citing judicial opinions that did not exist. ChatGPT had fabricated them, and when one of the attorneys asked the tool whether a cited case was real, it answered that the case existed and could be found in the standard legal databases. The court sanctioned the attorneys and their firm. Read against Table 1, the failure could have been intercepted at three points before filing. A Problem Framing probe (what was the

system asked to produce, and what did that request leave out?) would have surfaced prompts seeking supportive case law with no verification step. The Evidence Evaluation evidence, a trail showing each citation checked in a legal database, was absent, and checkably so, because that evidence is substrate-tied: anyone can look the cases up. Treating fluent, confident output as authoritative is automation bias, a failure documented in research on human use of automation long before generative AI (Parasuraman & Manzey, 2010). A Data and AI Governance review would have flagged unsupervised AI use in a court filing. Nobody needed to detect AI use to prevent this. They needed to look at the right surfaces.

Substitute a systems analyst for the attorney, and a requirements document, a data pipeline, or a security assessment for the filing, and the structure stays the same. Regulation is moving in the same direction. The European Union's Artificial Intelligence Act requires deployers of high-risk AI systems to assign human oversight to people "who have the necessary competence, training and authority" (European Parliament & Council of the European Union, 2024, Article 26(2)). Someone has to be that person, and many of them will be our graduates.

6. WHAT THIS ASKS OF US

6.1 Of Instructors

Keep the artifact and stop treating it as the evidence. Let students use AI to build the schema or the program and move the weight of the grade toward the process artifacts around it. Three design tests follow from Section 3. Can we re-check this evidence against something outside the student's own account? Could someone who was not paying attention have produced it honestly? Would it survive follow-up questions? An assessment element that fails all three is measuring the toolchain. Modest changes go a long way: a verification log attached to AI-assisted code, a decision log kept during a project, a short oral probe on a design choice selected at random.

We should be honest about the cost. Reading logs and holding conversations takes time that scoring a deliverable against a rubric did not, and it scales poorly to large sections. Oral probes raise equity questions for students with anxiety, students working in a second language, and students with accommodations, and they deserve the care we give any high-stakes format. None of this is a reason to keep grading artifacts that no longer carry information. It is a reason to design carefully and, as we argue next, to build tools.

6.2 Of Programs

Program-level assessment has the same exposure and has received far less attention. Assurance-of-learning and accreditation evidence commonly rests on rubric-scored student work products. If course artifacts no longer discriminate, program measures built on them do not either, and a program risks certifying its students' access to tools rather than its graduates' capability. Competency models tell us what to develop. They now need a companion account of how each competency is to be evidenced when AI is present.

This also presents an opportunity for our discipline. Capturing verification events, decision points, and framing choices as the work happens is an information systems design problem. The three properties and six surfaces can be read as requirements for that instrumentation: they specify what a learning platform, a development environment, or a project repository should be built to observe. IS and computing educators are unusually well placed to build and study such tools, beginning in our own courses.

6.3 Of Researchers

Our argument makes claims that could be wrong, and we would like to see them tested in IS and computing classrooms. First, as AI use intensifies, grades based on output artifacts and independent ratings based on process evidence should diverge in how they rank students. Second, process evidence should raise agreement among graders relative to artifact-only grading. Third, students' self-assessed competence should diverge from process-based ratings more under AI assistance than without it. A fourth, exploratory expectation is that ratings on the six surfaces will be correlated rather than independent, consistent with their compositional structure. If the first of these does not hold, our premise is wrong, and the field should know.

We also want to be clear about what we are not claiming. The six surfaces are a content specification, not a measurement instrument. Their support so far is content coverage: every incident examined could be classified on them. That is a weak test, since broadly defined categories applied by their author resist disconfirmation. Whether educators and practitioners recognize the surfaces, how they should be scored, and whether process evidence resists gaming once students know it is being graded are open questions. The public record also over-represents failures that end in litigation or in the press, so governance failures are probably over-weighted in our thinking.

7. CONCLUSION

The question of whether a student used AI has absorbed a great deal of our attention, and it has a short shelf life. The students now entering our programs will use AI in every course and every job, and the artifacts they submit will be good. What we can still observe is how they framed the problem, what they checked, what they understood about the setting, which risks they raised, how they directed the tools, and whether the result was worth having. Those are the things employers, courts, and regulators will hold them accountable for. We should teach them deliberately, and we should grade evidence that cannot be produced without them. We invite colleagues to test this framework, identify where it succeeds and where it falls short, and help establish what counts as credible evidence of learning when AI can produce the work.

8. REFERENCES

- Becker, B. A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., & Santos, E. A. (2023). Programming Is Hard – Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation. *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (pp. 500-506). New York, NY: ACM. <https://doi.org/10.1145/3545945.3569759>
- Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*, 140(2), 889-942. <https://doi.org/10.1093/qje/qjae044>
- Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling Theory: A Review and Assessment. *Journal of Management*, 37(1), 39-67. <https://doi.org/10.1177/0149206310388419>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Gimpel, H., Hall, K., Decker, S., Eymann, T., Gutheil, N., Lämmermann, L., Braig, N., Maedche, A., Röglinger, M., Ruiner, C., Schoch, M., Schoop, M., Urbach, N., & Vandirk, S. (2025). Using Generative AI in Higher Education: A Guide for Instructors. *Journal of Information Systems Education*, 36(3), 237-256. <https://doi.org/10.62273/QLLG7172>
- Jiang, Y., & Nakatani, K. (2025). Exploring Implementations of GenAI in Teaching IS Subjects and Student Perceptions. *Journal of Information Systems Education*, 36(2), 180-194. <https://doi.org/10.62273/WFHO1011>
- Leidig, P. M., & Salmela, H. (2022). The ACM/AIS IS2020 Competency Model for Undergraduate Programs in Information Systems: A Joint ACM/AIS Task Force Report. *Communications of the Association for Information Systems*, 50, 470-480. <https://doi.org/10.17705/1CAIS.05021>
- Long, D., & Magerko, B. (2020). What Is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-16). New York, NY: ACM. <https://doi.org/10.1145/3313831.3376727>
- Mata v. Avianca, Inc. (2023). 678 F. Supp. 3d 443 (S.D.N.Y. 2023).
- Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>

- Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A Method for Taxonomy Development and Its Application in Information Systems. *European Journal of Information Systems*, 22(3), 336-359. <https://doi.org/10.1057/ejis.2012.26>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3), 381-410. <https://doi.org/10.1177/0018720810376055>
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment. *Journal of University Teaching and Learning Practice*, 21(6). <https://doi.org/10.53761/q3azde36>
- Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B. N., & Savelka, J. (2023). The Robots Are Here: Navigating the Generative AI Revolution in Computing Education. *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education* (pp. 108-159). New York, NY: ACM. <https://doi.org/10.1145/3623762.3633499>
- Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, 87(3), 355-374. <https://doi.org/10.2307/1882010>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of Detection Tools for AI-Generated Text. *International Journal for Educational Integrity*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Zhang, X. (2025). Teaching Tip: Incorporating AI Tools Into Database Classes. *Journal of Information Systems Education*, 36(1), 37-52. <https://doi.org/10.62273/GKZI2477>

AUTHOR BIOGRAPHIES

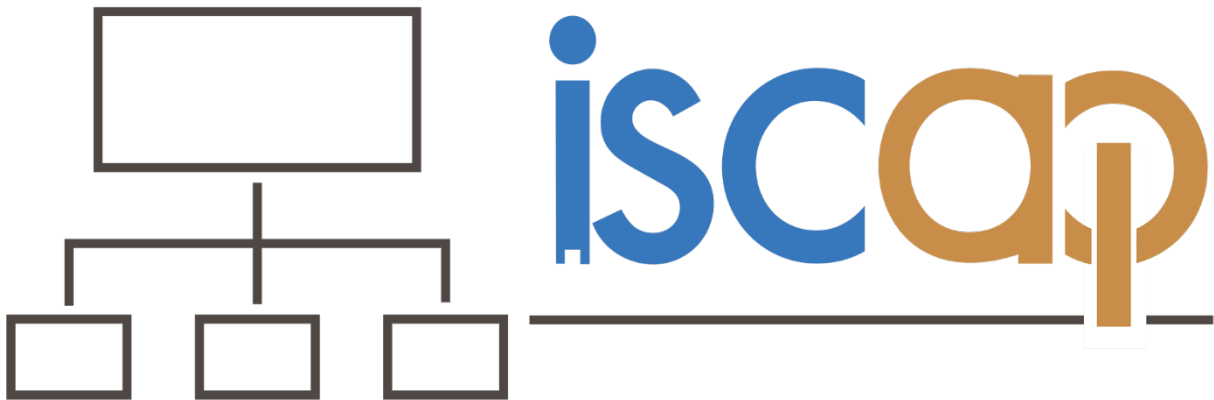
Timothy C. Smith is an Associate Professor of Instruction in the School of Information Systems at the University of South Florida's Muma College of Business, where he is the School's inaugural AI Fellow, and he serves as Technical Editor of the *Journal of Information Systems Education*. He teaches Python programming, machine learning, text analytics, generative AI, and enterprise systems development. His research examines generative AI, AI-enabled organizational change, and information systems education. Before joining academia, he spent more than 20 years in industry, including founding a software development company and holding engineering and executive roles in telecommunications. He received his Ph.D. in Management Information Systems from Carleton University.



Xihui "Paul" Zhang is a Professor of Computer Information Systems in the Sanders College of Business and Technology at the University of North Alabama in Florence, AL. He received his B.S. and M.S. degrees in Earth Sciences from Nanjing University (1993, 1996), Nanjing, China, and his M.S. and Ph.D. degrees in Management Information Systems from the University of Memphis (2004, 2009), Memphis, TN. His work has appeared in journals such as the *Journal of Strategic Information Systems*, *Information & Management*, *Journal of Database Management*, *Journal of Computer Information Systems*, and *ACM Transactions on Software Engineering and Methodology*. He is currently the Editor-in-Chief of *Journal of Information Systems Education*.



INFORMATION SYSTEMS & COMPUTING ACADEMIC PROFESSIONALS



STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the *Journal of Information Systems Education* have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

Copyright ©2026 by the Information Systems & Computing Academic Professionals, Inc. (ISCAP). Permission to make digital or hard copies of all or part of this journal for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial use. All copies must bear this notice and full citation. Permission from the Editor is required to post to servers, redistribute to lists, or utilize in a for-profit or commercial use. Permission requests should be sent to the Editor-in-Chief, *Journal of Information Systems Education*, editor@jise.org.

ISSN: 2574-3872 (Online) 1055-3096 (Print)