AI Assessment Futures: Creating Assignments That Withstand and/or Integrate AI

Robin Sutherland-Harris, PhD
Teaching Development Associate, Teaching Commons, York University

Today's Session

We will explore practical strategies for creating assessments that either minimize the risk of AI misuse or purposefully integrate AI tools to support transparent, ethical, and meaningful student learning.

YORK U



Overview

AI Assessment Futures: Creating Assignments That Withstand and/or Integrate AI

- State of the Field: Assessment Evolution
- Three Principles for Assessment
- Discussion: Case Scenarios

YORK U

How does GenAI work?

GenAI…

Identifies patterns:
- Not random! Identifies & selects statistically probably outputs
- No particular adherence to the truth – can “hallucinate”

Based on training data:
- Decides what is statistically probable based on training data
- Outputs reflect biases inherent in training data & society in general

To create original material:
- Technically original, but still bound by statistical probability
- The more you know about a topic, the more you can see where GenAI misses the mark

As it constantly changes:
- Agentic AI can now perform multi-step research & writing workflows autonomously – we are past AI just responding to single prompts
- AI-generated synthesis may now be mistaken for rigorous scholarship – this “illusion of understanding” goes beyond individual hallucinations

Consider cognitive offloading

Cognitive offloading is “the use of physical action [or a tool] to alter the information processing requirements of a task so as to reduce cognitive demand.”

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

See also:

- *Cecutti, L., Chemero, A., & Lee, S. W. S. (2021). Technology may change cognition without necessarily harming it. Nature Human Behaviour, 5(8), 973–975. https://doi.org/10.1038/s41562-021-01162-0*
- *Fan, Y., et al. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. British Journal of Educational Technology, 56(2), 489–530. https://doi.org/10.1111/bjet.13544*
- *Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. Quarterly Journal of Experimental Psychology, 74(8), 1477–1487. https://doi.org/10.1177/17470218211008060*

What do you think?

- GenAI has been around for almost four years.
  - In your experience, what has been the biggest shift in the assessment landscape that you would attribute to tech like ChatGPT, directly or indirectly?

State of the Field: Assessment Evolution in Light of AI



YORK U

Trends in AI and Assessment

1. Detection bias and equity concerns
2. Assessment design around AI
3. Critical thinking and productive friction

1. Detection bias and equity concerns

- **We are (still) unable to reliably use technology to guarantee the integrity of our assessments**
- What's happening: evidence continues to accumulate that AI-detection tools misclassify non-native English writing as AI-generated.
- Why it matters: early concerns about AI-detection are holding up under increased scrutiny, while a more nuanced picture is emerging.
- Course-level impact: Detection tools have a documented, patterned bias against multilingual and international students, making them unreliable guarantors of assessment validity.

2. Assessment design around AI

- **Deep assessment redesign to control for AI is increasingly inevitable.**
- What's happening: Responses to AI are increasingly moving away from asking "did a student use AI" and towards assessment redesign so that the question matters less.
- Why it matters: Protecting assessment integrity through policy/guidelines only creates an (exhausting) enforcement illusion.
- Course-level impact: The labour of meaningful assessment redesign falls primarily on faculty, but open opportunities for better teaching and deeper learning.

3. Critical thinking and productive friction

- **AI definitely (sometimes) harms thinking – but we can mitigate this.**
- What's happening: A more nuanced picture is emerging of how and where AI overreliance erodes critical thinking.
- Why it matters: More understanding of how this happens can help us determine how best to adapt our assessments to ensure productive friction.
- Course-level impact: Consider Furze's 5 areas for productive friction (expertise, evaluation, metacognition, stretch, feedback) and discipline/context-specific critical literacy in prompting and output evaluation.

Many models

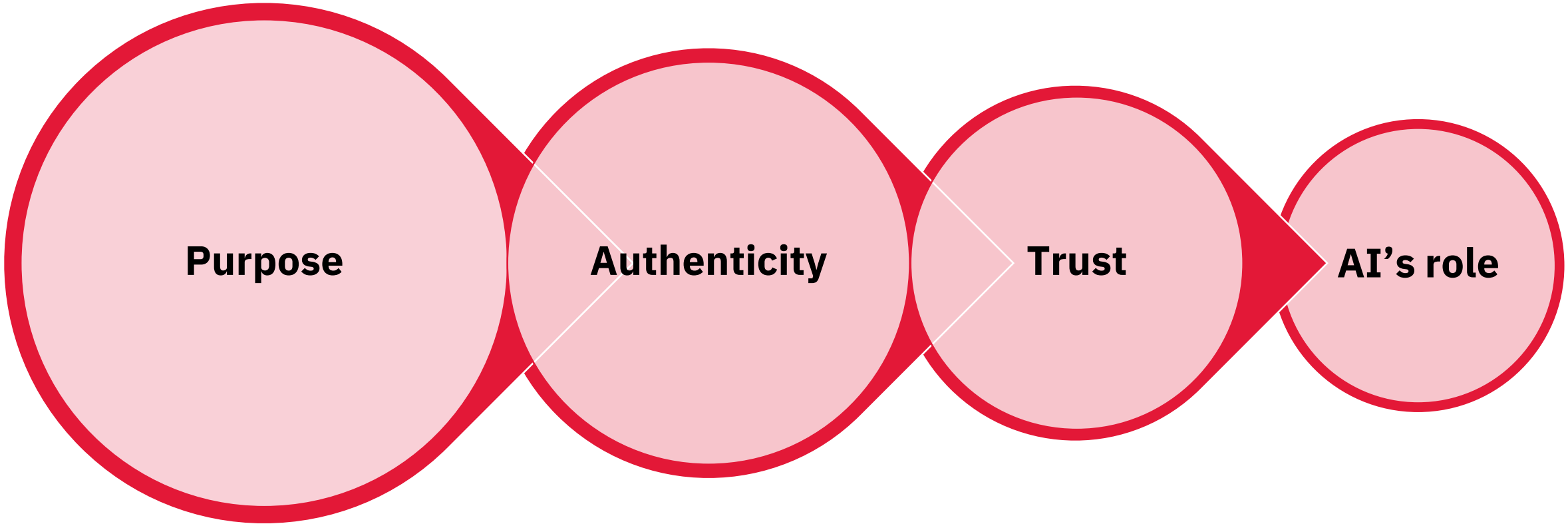
- Post-AI Bloom's Taxonomy
  - Hardman, D. P. (2024, October 3). A Post-AI Learning Taxonomy [Substack newsletter]. *Dr Phil's Newsletter*. https://drphilippahardman.substack.com/p/a-post-ai-learning-taxonomy
- Leon Furze's AI Assessment Scale
  - Perkins, M., Jasper, R., & Furze, L. (2025). Reimagining the Artificial Intelligence Assessment Scale: A refined framework for educational assessment. *Journal of University Teaching and Learning Practice*, *22*(7). https://doi.org/10.53761/rrm4y757
- Discursive vs. Structural Shifts
  - Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*, *50*(7), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964
- Two-Track Approach
  - *Program level assessment design and the two-lane approach—Teaching@Sydney*. (2024, September 12). https://educational-innovation.sydney.edu.au/teaching@sydney/program-level-assessment-two-lane/
- Black-Box Assessment
  - Winstone, N. E., Gravett, K., & Elkington, S. (2026). Black Box Assessment: Rethinking integrity and learning for a time of Generative AI. *Assessment & Evaluation in Higher Education*, *0*(0), 1–18. https://doi.org/10.1080/02602938.2026.2661365

Framework: Three waypoints



Three waypoints for your assessments

Before we make final decisions about AI in assessment, first consider:



YORK U

An example: A case-based analytical assessment

Consider a version of this assessment relevant to your discipline:

Students are given a real or realistic problem from the field. They analyse it, apply relevant disciplinary knowledge, and propose a reasoned response, submitted as a written report.

Purpose

Students are given a real or realistic problem from the field. They analyse it, apply relevant disciplinary knowledge, and propose a reasoned response, submitted as a written report.

- What is the intellectual work required of the student? What is this assessment meant to reveal to you about the learning that has taken place?

Authenticity

Students are given a real or realistic problem from the field. They analyse it, apply relevant disciplinary knowledge, and propose a reasoned response, submitted as a written report.

- What does it authentic engagement with the assessment look like? What does it mean for a student to genuinely do the thinking this assessment demands? What would superficial or substituted engagement look like by contrast?

Trust

Students are given a real or realistic problem from the field. They analyse it, apply relevant disciplinary knowledge, and propose a reasoned response, submitted as a written report.

- What does this assessment need in order to be trustworthy?

Purpose, authenticity, and trust: Scaling up

ASSESSMENT LEVEL

- *Purpose:* What is the intellectual work required of the student? What is this assessment meant to reveal to you about the learning that has taken place?
- *Authenticity:* What does it authentic engagement with the assessment look like? What does it mean for a student to genuinely do the thinking this assessment demands? What would superficial or substituted engagement look like by contrast?
- *Trust:* What does this assessment need in order to be trustworthy? How can you make trust-bearing elements stronger?

COURSE LEVEL

- *Purpose:* What does the full set of assessments across this course collectively evidence? Do they build toward something coherent, or are they a set of discrete tasks that don't add up?
- *Authenticity:* Does this course create genuine conditions for intellectual grappling? Or does the assessment sequence potentially allow students to move through without being challenged?
- *Trust:* What does the course's assessment design assume about how students are engaging with the material, with each other, with their own thinking? Are those assumptions supported by anything in the design, or are they just hopes?

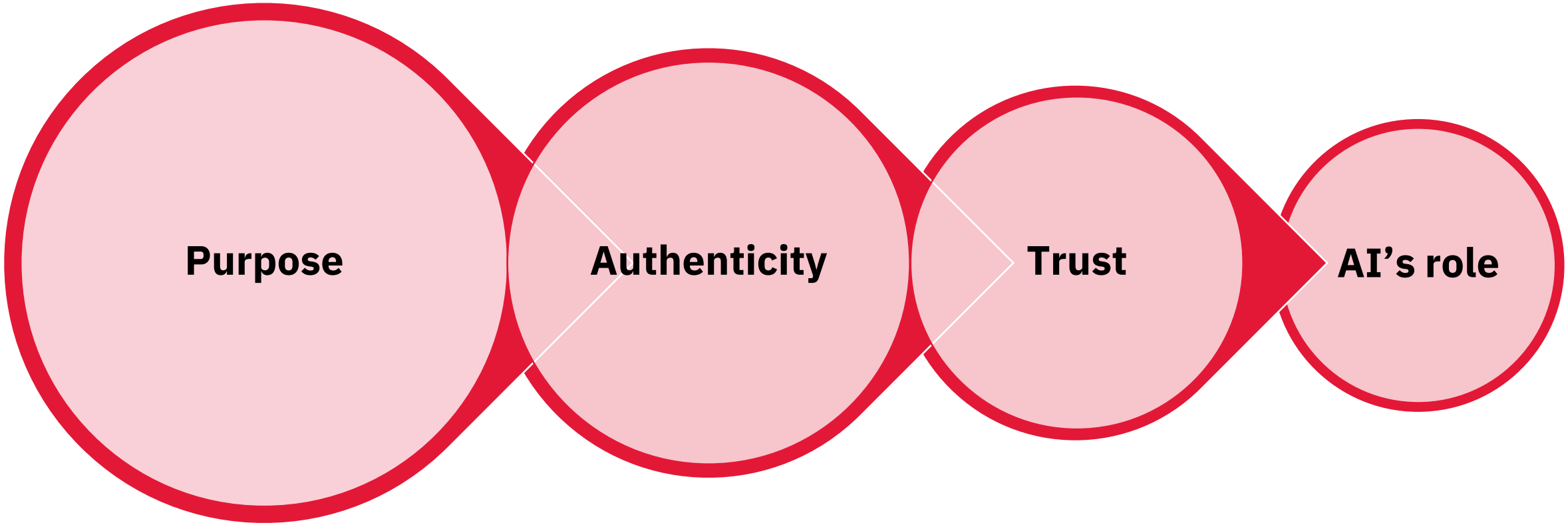
PROGRAM LEVEL

- *Purpose:* What capacity, disposition, or understanding does the program's assessment architecture actually develop in graduates? Is that visible in the design, or only assumed?
- *Authenticity:* Have graduating students had enough opportunities for authentic intellectual engagement — progressively demanding, across different contexts — to achieve program goals?
- *Trust:* What does the program assume students can do by graduation? Is there a coherent design basis for that trust, built progressively across the curriculum, or is it an assumption that has never been made explicit or tested?

YORK U

Three waypoints for your assessments

Before we make final decisions about AI, first consider:



YORK U

Case Scenarios



3 Vignettes

- Vignette 1: Workload and scale (scaffolding take-home essays)
- Vignette 2: Determining productive friction (STEM/foundational memorizing)
- Vignette 3: Equity (course project)

Each poses a scenario centred around assessment redesign that invites you to consider and discuss assessment purpose, authenticity, and trust in relation to some common challenges.

Thank you!

Robin Sutherland-Harris
Teaching Development Associate, Teaching Commons
York University
robinsh@yorku.ca

YORK U

Additional resources

- Agarwal, D., Naaman, M., & Vashistha, A. (2025). AI suggestions homogenize writing toward Western styles and diminish cultural nuances. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706598.3713564
- Ahangama, N. (2026). Designing assessments in the generative AI era: A tailored assessment framework for ICT tertiary education. *International Journal of Educational Technology in Higher Education, 23*(1), 9. https://doi.org/10.1186/s41239-026-00582-0
- Ajjawi, R., Tai, J., Bearman, M., Boud, D., Dawson, P., & Jorre de St Jorre, T. (2024). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education, 49*(4), 499–510. https://doi.org/10.1080/02602938.2023.2271193
- Basu, P., Zhang, Y., & Raheja, V. (2025). *BAID: A benchmark for bias assessment of AI detectors* (arXiv:2512.11505) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.11505
- Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*. Advance online publication. https://doi.org/10.1080/02602938.2025.2503964
- Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 49*(7), 1005–1016. https://doi.org/10.1080/02602938.2024.2386662
- Fan, Y., et al. (2025). Beware of metacognitive laziness. *British Journal of Educational Technology*. Advance online publication. https://doi.org/10.1111/bjet.13544
- Furze, L. (2026, March 22). *Resistance as a framework for combating cognitive offload*. https://leonfurze.com/2026/03/22/resistance-as-a-framework-for-combating-cognitive-offload/
- Higher Education Policy Institute. (2026a, July 20). *Catching the wrong students: AI detection, international students and the fairness crisis in UK universities*. https://www.hepi.ac.uk/2026/07/20/catching-the-wrong-students-ai-detection-international-students-and-the-fairness-crisis-in-uk-universities/

YORK U

Additional resources

- Higher Education Policy Institute. (2026b, July 14). *Verifiable judgment: What AI actually demands of universities*. https://www.hepi.ac.uk/2026/07/14/verifiable-judgment-what-ai-actually-demands-of-universities/
- Hossain, S. (2026). *Four types of LLM reliance and their predictors among undergraduate writers: A mixed-methods study at a minority-serving R1 university* (arXiv:2606.28749) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2606.28749
- Jiang, Y., & Chen, H. (2026). AI-based cognitive offloading by students in higher education in China. *Computers in Human Behavior: Artificial Humans, 8*, Article 100327. https://doi.org/10.1016/j.chbah.2026.100327
- Kosmyna, N., et al. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing tasks* (arXiv:2506.08872) [Preprint]. arXiv. https://arxiv.org/abs/2506.08872
- Laidlaw, P. (2026). Supporting assessment redesign in the age of AI: A critical practice-based framework for equitable change. *Teaching in Higher Education*. Advance online publication. https://doi.org/10.1080/13562517.2026.2670362
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*. https://doi.org/10.1016/j.patter.2023.100779
- Lodge, J. M., Howard, S., Bearman, M., Dawson, P., & Associates. (2023). *Assessment reform for the age of artificial intelligence*. Tertiary Education Quality and Standards Agency. https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assessment-reform-age-artificial-intelligence
- Misiejuk, K., López-Pernas, S., Kaliisa, R., & Saqr, M. (2026). Cognitive offloading in student–AI collaboration: A longitudinal analysis of prompting strategies. *Computers in Human Behavior Reports, 22*, Article 101130. https://doi.org/10.1016/j.chbr.2026.101130

YORK U

Additional resources

- Mollick, E. (2024, August 30). *Post-apocalyptic education*. One Useful Thing. https://www.oneusefulthing.org/p/post-apocalyptic-education
- Ransome, E. (2026, February 6). *What generative AI reveals about assessment reform in higher education*. Higher Education Policy Institute. https://www.hepi.ac.uk/2026/02/06/what-generative-ai-reveals-about-assessment-reform-in-higher-education/
- Roe, J., Perkins, M., & Giray, L. (2026). Assessment twins: A protocol for AI-vulnerable summative assessment. *Journal of Applied Learning and Teaching, 9*(2). https://doi.org/10.37074/jalt.2026.9.2.3
- Roy, B. K., & Jony, M. S. (2026). Generative AI and critical thinking in higher education: Students' narratives of cognitive offloading. *Journal of Interdisciplinary Studies in Education, 15*(3), 263–282. https://doi.org/10.32674/jey59q25
- Salamonson, Y., et al. (2026). Refining assessment design and practice in nursing and midwifery education in the era of generative AI: A discussion paper. *Journal of Clinical Nursing*. Advance online publication. https://doi.org/10.1111/jocn.70484
- Seung, Y., & Basham, J. D. (2026). Cognitive offloading in the age of generative AI: What does it mean for students with learning disabilities? *Learning Disability Quarterly, 49*(3), 121–133. https://doi.org/10.1177/07319487261439132
- Stowe, K., Afanaseva, S., Raimundo, R., Sun, Y., & Patil, K. (2026). *Identifying bias in machine-generated text detection* (arXiv:2512.09292) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.09292
- Times Higher Education. (n.d.). *AI policies are repeating every mistake of plagiarism policies* [Opinion]. https://www.timeshighereducation.com/opinion/ai-policies-are-repeating-every-mistake-plagiarism-policies
- Watkins, M. (2024, July 12). *The rhetoric of purpose: Conversations for our AI era*. Rhetorica. https://marcwatkins.substack.com/p/the-rhetoric-of-purpose-conversations
- Winstone, N. E., Gravett, K., & Elkington, S. (2026). Black box assessment: Rethinking integrity and learning for a time of generative AI. *Assessment & Evaluation in Higher Education*. Advance online publication. https://doi.org/10.1080/02602938.2026.2661365