

Teaching with Artificial Intelligence

MODULE 3

Assessment Stress Testing and the AI Assessment Scale

Faculty Development Reading

*An assessment is defensible when you can say what it measures,
what could fool it, and why you still trust it.*

— the course stance, in one sentence

June 2026

Module Purpose

Module 2 gave you a diagnosis: somewhere in your assignment, essential cognitive work is exposed to an AI bypass. Module 3 turns that diagnosis into assessment judgment. The module's central distinction — between **learning production** and **learning measurement** — explains why so many AI conversations go in circles: faculty are defending assignments that were quietly doing two jobs at once, and AI breaks each job differently. Its central tool, the **AI Assessment Scale (AIAS)** developed by Furze, Perkins, and Roe, replaces the blunt instruments of “AI allowed” and “AI banned” with a five-level vocabulary precise enough to apply to individual components of a single assignment (Perkins, Furze, Roe, & MacVaugh, 2024).

The portfolio artifacts are an **Assessment Stress Test** — a structured failure analysis of one assignment — and an **AIAS Decision Memo** (300–450 words) that a colleague could read, challenge, and act on without further clarification. The major decision you are preparing to make: *which AIAS level, for which components of this assignment, justified against the strongest counterargument you can articulate?* Note the last clause. This module does not ask you to label an assignment and move on; it asks you to argue against yourself first.

Learning Objectives

After completing this reading and the Module 3 activities, you will be able to:

1. Distinguish learning-production activities from learning-measurement activities (and both from logistics) in a course plan, in writing a colleague could verify.
2. Apply the AI Assessment Scale to one specific assignment — at the component level where levels differ — and justify the selected level using the learning outcome from your Friction Audit.
3. Conduct a four-question Assessment Stress Test identifying AI capability, invisible learning, process-evidence options, and equity/access risks, with at least one named scenario per question.
4. Choose among the five decision endpoints (Keep, Revise, Split, Replace, Integrate AI) and defend the choice.
5. Draft an AIAS Decision Memo that states your decision, the evidence behind it, and the strongest counterargument to it.

Before You Read: A 60-Second Retrieval Primer

YOUR 60-SECOND RESPONSE — write before you read

Name one assignment in your course. Is it teaching something, measuring something, or both? If both — where exactly is the line? Sixty seconds, one or two sentences. Most faculty discover the line is blurrier than they assumed; that blur is this module's raw material.

Core Reading

Why “AI allowed” and “AI banned” are both too blunt

Course-wide AI rules fail for a structural reason: a course is not one cognitive event. A blanket “no AI” rule may be coherent in a foundational course where independent schema-building is the whole point, and unnecessarily restrictive in an advanced course where evaluating AI output is part of professional practice. A blanket “AI welcome” rule may fit an applied studio and quietly hollow out an introductory course. Even within one assignment, planning, drafting, revision, and final demonstration ask students to do different intellectual work — and may need different rules (Module 3 lectures). The unit of decision is not the course, and often not even the assignment. It is the *component*. That is the precision the AIAS vocabulary exists to provide.

Production, measurement, and logistics

Before any AIAS level can be chosen, an older confusion has to be untangled. Every assignment component does one of three jobs. **Production** builds skill through effort, feedback, and revision — it is practice. **Measurement** generates evidence that a student can meet an outcome — it is proof. **Logistics** are process steps with neither pedagogical nor evidentiary function: file naming, submission format, the header block. The jobs are routinely conflated: a draft asked to be both practice and final proof; a discussion post designed for processing but graded for polish; a peer review treated as both a learning move and an assessment artifact.

AI is what makes the conflation expensive. When AI is permitted for production but the assignment is graded as measurement, the grade no longer reflects the learning. When AI is banned for measurement but the production that should have preceded it was outsourced, the measurement faithfully reveals the production gap — and the student fails a test that was really assessing the homework. The Stress Test worksheet operationalizes the untangling as a **three-color annotation**: mark each component production (green), measurement (red), or logistics (gray). Components you cannot color are usually the problem.

The equity stakes are larger than they first appear. When production and measurement are conflated, students who arrived with schemas — from stronger schools, from homes with academic English, from prior degrees — pass measurements that secretly assess production they demonstrated elsewhere. Students without those schemas fail the same measurements for reasons the syllabus never named. AI intensifies the split: equipped students use it to amplify their schemas; unequipped students use it to substitute for schemas they were never given a fair chance to build (Module 3 lectures).

The AI Assessment Scale: vocabulary, not a rulebook

The AIAS (Perkins, Furze, Roe, & MacVaugh, 2024) names five levels of AI involvement. The course’s framing matters as much as the scale itself: AIAS is a *vocabulary* for making expectations visible, not a rulebook to be pasted across a course. Each level is best read as a **contract** that aligns three things at once: what the student does, what AI does, and what evidence of learning is required.

Level	What the student does	What AI does	Evidence and documentation	Typical equity concern
1 — No AI	Works independently; the product demonstrates independent capability.	Nothing.	Independent product; no documentation needed.	Low tool-access risk; needs a stated learning rationale so it does not read as merely restrictive.
2 — AI Planning	Uses AI to brainstorm or outline; produces the assessed work personally.	Supports planning only.	Student-produced work plus a brief planning note (one sentence).	Minimal; keep documentation light.
3 — AI Collaboration	Works with AI during the task; retains intellectual responsibility.	Contributes during the work, documented.	Final product plus a decision record: suggestions accepted, rejected, revised.	Assumes tool access; require a free-tier path and clear documentation norms.
4 — Full AI	Curates, evaluates, integrates; explains what changed and why.	May be the primary producer.	Stronger evidence of curation and judgment; the polished product alone reveals little.	Highest access stakes; an access plan is required, plus a bias acknowledgment where relevant.
5 — AI Exploration / Critique	Studies and critiques AI behavior itself — method, evidence, limits, effects.	Is the object of study.	The critique, not the AI output, is the submission.	Requires enough schema to critique meaningfully; scaffold for novices.

Table 1. The five AIAS levels as contracts (after Perkins, Furze, Roe, & MacVaugh, 2024, as adapted in the Module 3 Quick Reference card). Each level aligns student work, AI's role, and the evidence required.

Documentation should be proportionate to the contract: Level 2 may need one planning sentence; Level 3 needs a decision record; Level 4 needs real evidence of curation, because the polished product no longer reveals the student's contribution. And the scale carries a paradigm with it: Eaton's (2023) **post-plagiarism** argument that integrity practice is shifting from proving the absence of AI toward documenting contextually appropriate human–AI collaboration. AIAS operationalizes that shift — which also means a Level 1 choice is not a refusal of the paradigm but a move within it: “no AI here, because this is where independent capability is demonstrated” is documented, contextual, and defensible.

One course, several levels — coherently

The course's running example is an introductory statistics course. The early probability quiz is Level 1: the instructor needs evidence of independent method selection. Later homework is Level 2: AI may help plan an approach, but students produce the solutions. The final data-analysis project is Level 3: collaboration with documentation, because documented judgment is part of the outcome. A unit on detecting fabricated statistical claims is Level 5: AI output is the object of critique. Same course, same students, four contracts — and the variation is not inconsistency. It is constructive alignment (Biggs, 1996): the AI rule changes because the learning goal and the evidence change. One explanatory sentence per assignment is usually enough: *this quiz is Level 1 because I need evidence of your independent method selection; this project is Level 3 because documented collaboration is part of the outcome.*

COMMON TRAP — the permission sticker

The most common misuse of AIAS is treating it as a permission sticker: choose a level that matches your comfort with AI, paste it on the syllabus, done. Two tells give it away. First, the level is identical for every task — which means it was chosen for the instructor's identity, not the assignments' outcomes. Second, the rationale mentions the tool but not the learning ("I selected Level 1 because AI undermines academic integrity" names no outcome, no evidence, no component).

The repair is the question the Stress Test forces: *what evidence of which capability does this component need to produce?* Choose the level that protects that evidence. A defensive Level 1 and a fashionable Level 3 fail the same way — by skipping the question.

The Stress Test: a failure analysis, not a checklist

With the vocabulary in hand, the module's tougher protocol begins. You do not simply label your assignment; you **stress-test** it — a structured failure analysis in the engineering sense, run before the failure happens in a live classroom. Four questions, each demanding at least one named, specific scenario:

1. **AI capability.** What could a current, general-purpose AI tool produce for this assignment — concretely? If you have AI access, test it; if not, describe your assumption and date it (capability claims in this reading: June 2026). Vague dread and vague dismissal fail this question equally.
2. **Invisible learning.** If the submitted product looked good, what learning would remain invisible? Name the capability the product alone cannot show — your Module 2 essential operation usually reappears here.
3. **Process evidence.** What evidence would restore confidence that the student did the learning? The menu includes an annotated draft, a short decision memo, a brief recorded oral defense, an LMS quiz with short-answer follow-up, a prompt log with reflection, or a student-versus-AI comparison with critique. Pick the lightest form that reveals a real decision — Module 4 enforces a rule of one.

4. **Equity and access risk.** Does the design require paid tools? Expose private data? Are multilingual students or students with disabilities disadvantaged? Is the grading load feasible at your actual enrollment, with your actual staffing?

The Stress Test ends at a decision with five endpoints: **Keep** the assignment as is (now with stated reasons); **Revise** it; **Split** it — separate the production components from the measurement components and rule them differently; **Replace** it with a different task; or **Integrate AI explicitly** and assess the judgment the integration demands. Faculty who skip the test tend to choose levels by preference or fear. Faculty who run it tend to discover the decision was more specific — and more defensible — than either.

EQUITY AND ACCESS CHECK

Stress Test question 4 deserves its own discipline, because its failures are quiet. A Level 3 or 4 choice assumes access to tools: if your institution does not provide AI, the course floor applies — every required use must be completeable on a free tier, on a phone, on a slow connection. Process evidence has its own equity profile: a recorded oral defense that steadies one student terrifies another and burdens the student working two jobs; offer an equivalent written or asynchronous form.

And weigh the grading load honestly: a process portfolio that is feasible for a seminar of twelve is a different proposition for an adjunct teaching three sections of sixty. An assessment design that only works with resources your colleagues do not have is not a design; it is a memo from a smaller class.

The Decision Memo — and the counterargument that makes it defensible

The module's written artifact is deliberately a *memo*, not a reflection: a professional genre addressed to a colleague who could understand, challenge, and act on it. The worksheet template runs five moves — the assignment, the AIAS level selected, the Stress Test summary, the decision, and what will change in Module 4 — in 300–450 words. The hardest and most valuable sentence states the **strongest counterargument** to your own choice, fairly, followed by what evidence would change your mind. If you chose Level 1, the strongest counterargument may be that you are measuring stamina under prohibition rather than capability, or that unequal underground AI use will now masquerade as differences in talent. If you chose Level 3, it may be that documentation invites performative logs while the actual thinking remains outsourced. Writing the objection yourself is the move that separates a defensible decision from a preference with paperwork — and it is the exact cognitive work a skeptical colleague would have done for you in a department meeting.

The module supplies a bounded **AI critical-friend protocol** for this step: a fixed prompt instructing the AI to ask you three Socratic questions about your memo — where you claim something unjustified, what a skeptical colleague would find unconvincing, which sentence a student would misunderstand — and explicitly forbidding it to rewrite or edit. The boundary is the pedagogy: the memo is your reasoning, and the protocol surfaces weaknesses without substituting for the schema-building (Module 3 critical-friend prompt; a paper variant exists for participants without AI access, because the questions are the value, not the model).

Transparent and authentic — the bridge to Module 4

The strongest redesigns do not merely make cheating harder; they make purpose, task, and criteria clearer — and they situate performance in a context with disciplinary meaning. That is the TILT transparency framework (Winkelmes) plus authentic assessment (Wiggins, 1998), and Module 4 builds your assignment around both. The preview matters here because the Stress Test often reveals that the assignment’s real problem was never AI: it was that students could not see what the task was for, what counted as quality, or what situation it rehearsed. “Real world” decoration does not fix that; a named audience, genre, and constraint a practitioner of your discipline would recognize might. Your Decision Memo’s final field — *what I will change in Module 4* — is where that diagnosis lands.

PAUSE AND APPLY — run one question now

Take the assignment from your Friction Audit and run Stress Test question 2 only: *if the product looked good, what learning would remain invisible?* Write the answer as one sentence beginning “The product alone cannot show that the student can...”

If you cannot finish the sentence, the assignment may be measuring nothing the product doesn’t already give away — which is itself a finding, and a common one. Bring the sentence to the worksheet; it becomes the spine of your memo.

Connection to the Module Artifact

CONNECTION TO YOUR PORTFOLIO ARTIFACT — Assessment Stress Test + AIAS Decision Memo

The Stress Test worksheet carries your Module 2 work forward: the same assignment, the outcome restated in behavioral language, the three-color annotation, a provisional AIAS level (whole assignment, then components where levels differ), the four questions, the five-endpoint decision, and the 300–450-word memo with its counterargument.

The memo becomes the design brief for Module 4’s redesign — so write it to the colleague who will challenge it, not to the steward who will read it kindly. The samples library (three memos across disciplines, randomized) calibrates your eye before you self-score against the rubric.

Before you submit, confirm:

- ☐ My learning outcome is stated in behavioral language — verbs like distinguish, apply, critique, not “understand.”
- ☐ Every component carries a color: production, measurement, or logistics — and conflated components are named, not averaged.
- ☐ My AIAS level is assigned at the component level where components genuinely differ, with a one-sentence learning rationale per level.

- All four Stress Test questions have at least one named, specific scenario — including a dated AI-capability claim (June 2026).
- My decision endpoint (Keep / Revise / Split / Replace / Integrate) follows from the test, and I could defend it in a department meeting.
- My memo is 300–450 words, states the strongest counterargument fairly, and names the evidence that would change my mind.

Key Terms

Term	What it means in this course
Learning production	Assignment work that builds skill through effort, feedback, and revision — practice, not proof.
Learning measurement	Assignment work that generates evidence a student can meet an outcome — proof, not practice.
Logistics	Process steps with neither pedagogical nor evidentiary function (file naming, formatting, submission mechanics).
Three-color annotation	Marking each component green (production), red (measurement), or gray (logistics) to expose confluents.
AI Assessment Scale (AIAS)	A five-level vocabulary for AI involvement in assessment — No AI; AI Planning; AI Collaboration; Full AI; AI Exploration/Critique (Perkins, Furze, Roe, & MacVaugh, 2024).
Contract (AIAS sense)	What a level really specifies: what the student does, what AI does, and what evidence is required — together.
Component-level decision	Assigning AIAS levels to parts of an assignment rather than pasting one level across a course.
Assessment Stress Test	A four-question failure analysis: AI capability, invisible learning, process evidence, equity/access risk.
Process evidence	Artifacts that make thinking visible — annotated drafts, decision memos, brief oral defenses — chosen at the lightest sufficient weight.
Decision endpoints	The five outcomes of a Stress Test: Keep, Revise, Split, Replace, Integrate AI explicitly.
AIAS Decision Memo	A 300–450-word professional memo stating the level, the evidence, the decision, and the strongest counterargument.
Post-plagiarism	Eaton’s (2023) paradigm: integrity through documented, contextually appropriate human–AI collaboration rather than proof of AI’s absence.

Brief Applied Example

A nursing ethics instructor splits an assignment. Professor Tran's signature assessment is a case analysis: students receive a patient scenario and write a recommendation memo to a program director. Her Friction Audit named the essential operation as *apply* — using ethical frameworks to reach a defensible recommendation. The Stress Test is sobering on question 1: a current general-purpose model produces a competent framework-citing recommendation for her stock cases in under a minute (claim dated June 2026). Question 2 names what would be invisible: whether the student can weigh the frameworks against the messy particulars — the family's conflict, the resource constraint — rather than reciting them. Question 4 flags two risks: her cases circulate among student cohorts, and her sixty-student sections make a per-student oral exam infeasible.

Her endpoint is **Split**. The written memo becomes Level 2: AI may be used to brainstorm framework-relevant questions (documented in one sentence), but the recommendation and reasoning are the student's — the memo remains the production-rich practice space. The measurement moves to a fifteen-minute, in-class component: students receive one *new* twist on the case (a changed family decision, a new constraint) and handwrite a half-page update to their own recommendation — pure apply, no AI, feasible to grade with a two-row rubric. Her memo's counterargument is honest: a skeptical colleague would say the handwritten format disadvantages multilingual students and students with disabilities, and she answers it in the design — extended time per existing accommodations, and the twist response graded for reasoning, never prose polish. What would change her mind: two terms of twist responses showing no correlation with memo quality, which would tell her the split is measuring composure, not judgment. The whole memo runs 410 words, and her chair could act on it tomorrow.

End-of-Reading Reflection

1. Return to your primer response. Redraw the line: which parts of the assignment you named are production, which are measurement, and which were logistics wearing a rubric? What changed?
2. For your own course or Synthetic Course Packet: which single component most needs a different AIAS level than the rest of the assignment — and what is the one-sentence learning rationale you would give students?
3. Write, in two sentences, the strongest counterargument to the AIAS level you are currently leaning toward. If you cannot write it, whose objection do you most need to hear before Module 4?

Sources and Further Reading

The module materials cited throughout this reading — the Module 3 lectures, the Module 3 Participant Workbook, the AIAS Quick Reference card, the Critical-Friend prompt card, and the Decision Memo samples library — are provided inside the course.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3).

Eaton, S. E. (2023). Postplagiarism: Transdisciplinary ethics and integrity in the age of artificial intelligence and neurotechnology. *International Journal for Educational Integrity*, 19.

Harvard University. (n.d.). *AI teaching resources*. <https://www.harvard.edu/ai/teaching-resources/>

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 & Learning Practice*, 21(6).

Stanford Teaching Commons. (n.d.). *Artificial intelligence teaching guide*. Stanford University. <https://teachingcommons.stanford.edu/teaching-guides/artificial-intelligence-teaching-guide>

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Winkelmes, M.-A. (n.d.). *Transparency in Learning and Teaching (TILT Higher Ed)*. <https://tilthighered.com>

Suggested optional extension: Furze's blog commentary accompanying the AIAS (leonfurze.com) shows the scale's evolution and common implementation mistakes; the Stanford guide's assessment workshop pairs well with the Stress Test's question 3 menu.