



The Credential Mirage in AI-Era Higher Education: A Counterfactual Competence Audit for Trustworthy Degrees

Kwan Hong TAN

Singapore University of Social Sciences, Singapore

DOI: <https://doi.org/10.68050/JAMS.2026.869>

Received: 14 October 2026; Accepted: 20 October 2026; Published: 26 October 2026

ABSTRACT

Generative artificial intelligence (AI) can raise the quality of assessed student outputs without necessarily establishing the independent or transferable competence implied by a degree. This article calls the potential gap between a warranted capability claim and an overly broad credential inference the credential mirage. Drawing on argument-based validity, authentic and programmatic assessment, cognitive offloading, evaluative judgment, and the AI Assessment Scale, it develops the Counterfactual Competence Audit (CCA) through analytical design science. CCA is a proposed program-level claim-validation protocol, not an alternative AI-permission scale or an empirically validated instrument. It combines bounded independent performance (Mode A), accountable AI augmentation (Mode B), and delayed changed-context transfer (Mode C), and links six design principles to evidence sources, four rubric anchors, claim-specific decision rules, and five falsifiable propositions. A hypothetical MBA capstone illustrates how strong AI-assisted work can coexist with insufficient evidence for an indispensable independent outcome, without imputing misconduct. Local standard setting, accessible assessment, calibrated moderation, privacy protection, and conditional or deferred decisions are essential safeguards. The contribution is an explicit chain from assessed performances to degree claims, including where that inference must be withheld. Prospective, cross-disciplinary validation remains necessary; no participant study or observed treatment effect is claimed.

Keywords: higher education; generative artificial intelligence; assessment validity; graduate competence; learning transfer; counterfactual competence audit

1. INTRODUCTION

A university may award an excellent mark to a sophisticated report, code repository, or strategic analysis even when the student could not independently diagnose a closely related problem, defend a decision, or recognize an invented citation. This possibility does not establish that AI causes incompetence. It establishes a measurement problem: the observed excellence of an artifact and the competence of its nominal author are no longer reliably interchangeable. As generative AI becomes embedded in research, writing, programming, and professional work, this distinction threatens a central promise of higher education, namely that qualifications communicate defensible capabilities beyond the tasks completed for credit.

Evidence on learning with AI is heterogeneous. In a randomized study of a university physics course, Kestin et al. (2025) found that a specially designed AI tutor supported stronger immediate learning than an active-learning comparison. By contrast, Bastani et al. (2025), studying high-school mathematics, found that access to a general-purpose AI system without suitable guardrails could undermine performance when assistance was withdrawn. These studies differ in populations, tools, tasks, and outcomes; they do not justify a simple claim that AI either improves or destroys higher education. They do show why assessment must specify which capability, under which conditions, is being certified. More broadly, generative AI has demonstrated gains in professional task productivity (Noy & Zhang, 2023), but professional output efficiency does not itself establish learning or durable transfer.

The practical temptation is to choose between permissive AI usage and prohibitive examination conditions.



Both extremes are insufficient. Blanket prohibition neglects the competence to supervise AI in occupations where AI is a normal instrument. Unqualified permissiveness risks confusing artifact production with mastery. The Artificial Intelligence Assessment Scale (AIAS) appropriately clarifies degrees of permitted AI involvement (Perkins et al., 2024; Perkins et al., 2025), while educational scholarship stresses evaluative judgment about AI outputs and processes (Bearman et al., 2024). Yet a further question remains: what evidence is sufficient to justify the particular inferential claim made by an awarded degree?

This article proposes the Counterfactual Competence Audit (CCA) as a design response. Its unit is not the detector, the individual essay, or an abstract policy on prompts. Its unit is the evidence chain supporting a credential claim. “Counterfactual” refers to deliberately sampling plausible performance conditions, including reduced assistance and altered contexts, to test whether conclusions drawn from a high-quality artifact survive changes in support and task demands. Because human capabilities cannot literally be observed in every alternative world, the audit produces bounded evidence rather than a causal certainty or a universal measure of intelligence.

The contribution is differentiated from three adjacent approaches. It does not replace AIAS, which indicates levels of permitted AI use; it supplements such disclosure with validity evidence. It differs from AI tutoring frameworks by assessing what is credentialed across modes rather than optimizing a specific instructional intervention. It also extends beyond resource-constrained learning assurance (Tan, 2026a) by examining the interpretation and governance of qualification claims at program level. The central research question is: How can a university make credible, equitable, and future-facing claims about graduate competence when assessed outputs may be co-produced with generative AI?

2. THEORETICAL BACKGROUND AND CONCEPTUAL GAP

2.1. Credential validity as an argument about inference

Assessment validity is not an intrinsic property of a mark or test format; it concerns the defensibility of interpretations and uses. Kane (2013) frames validation as an argument linking observed performance to interpretations and consequences. Applied to an AI-mediated degree, the inferential chain is: assessed evidence, defensible scoring, generalization across tasks, extrapolation to future contexts, and a justified credential decision. AI complicates at least the scoring-to-generalization and generalization-to-extrapolation steps. A polished output may reflect tool capability, access to premium models, collaborator intervention, or effective student orchestration in varying proportions. Without deliberate evidence collection, the institution cannot tell which inference is warranted.

The source of assessment failure is therefore not merely undetected AI use. An assessment may transparently authorize AI and still support invalid inferences if its published learning outcomes promise unaided reasoning that was never evaluated. Conversely, an unassisted timed task may be internally reliable yet poorly aligned with professional capabilities requiring tool selection, supervisory judgment, and collaborative problem solving. The appropriate target is construct validity: the match between the capability claimed and the performance sampled (Kane, 2013; Boud & Falchikov, 2006).

2.2. Cognitive offloading, learning and performance

Cognitive offloading denotes shifting part of a cognitive task to external tools or actions (Risko & Gilbert, 2016). Offloading can be rational and beneficial, such as freeing attention for higher-order judgment; however, an externalized solution does not necessarily produce durable internal representations. Earlier work on distributed information access demonstrated that people may remember where information can be found rather than its content (Sparrow et al., 2011). Retrieval practice research shows that independently generating answers can support later retention and transfer (Roediger & Karpicke, 2006; Butler, 2010). Taken together, these findings motivate sampling performance after support changes, not an indiscriminate expectation that everything must be memorized.

The distinction between performance and learning matters especially for fluent AI systems. The same immediate output can arise from deep understanding, effective delegation, superficial prompting, or chance. Moreover, a generative model that provides hints, questions, and formative feedback may strengthen understanding, whereas a model providing complete answers may bypass productive cognitive activity (Bastani et al., 2025; Kestin et al., 2025). A credible audit must separate these mechanisms rather than infer them from tool use alone.

2.3. Evaluative judgment, transfer and the purpose of a degree

Bearman et al. (2024) argue for cultivating evaluative judgment of AI products and processes. This implies that the competent graduate is neither a tool-free problem solver in every situation nor an uncritical operator of a black box. Competence is multidimensional: some elements must be available independently, while others are demonstrated through reasoned augmentation and accountability. Transfer adds a time and context dimension: can the learner mobilize principles when superficial features change, when the tool becomes unavailable, or when its answer is misleading? Long-term, future-oriented assessment theory likewise emphasizes the capabilities learners take into situations beyond the course (Boud & Falchikov, 2006).

Prior work by Tan (2025a) argues that AI-augmented postgraduate education should develop complementary human and technological strengths; Tan (2025b) shifts attention from degree possession to transferable skills in doctoral education. Tan (2026b) discusses structured reflective dialogue as a means of exercising evaluative and inferential skills among adult learners. These works motivate the program-level question advanced here but do not establish the validity of the CCA artifact. Ethical concerns about pedagogy further require that a credential audit protect inclusion, dignity, and reasonable accommodations rather than equate ability with performance in a single stressful modality (Tan, 2025c).

2.4. Positioning the CCA against adjacent assessment traditions

CCA does not claim to invent validity argumentation or evidence aggregation. Messick (1995) establishes that interpretations and consequences are central to validity; Kane (2013) specifies claims and warrants that must survive scrutiny. Authentic assessment asks whether activities and participation resemble valued practices, including future professional settings (Ajjawi et al., 2024). Programmatic assessment accumulates information across observations and separates lower-stakes feedback from consequential decisions (van der Vleuten et al., 2012). AIAS makes permissible AI participation explicit (Perkins et al., 2024, 2025). These approaches solve different, valuable parts of the problem.

The proposed incremental contribution is a credential-level contrast protocol. It treats support conditions and a delayed context change as deliberately varied sources of evidence, maps each source to a narrow claim, and supplies a documented adjudication rule when observations conflict. Its testable novelty is therefore not that three tasks always outperform one, but whether claim-specific multimode evidence adds useful validity beyond well-designed authentic, programmatic and AI-governed assessment already in place. Table 1 makes this boundary explicit.

Table 1. Relationship between established approaches and the proposed CCA

Approach	What it already contributes	What remains undecided	CCA-specific addition
Argument-based validity (Messick; Kane)	Explains why interpretation, extrapolation and consequences require evidence.	Which AI-affected observations should warrant a particular graduate claim.	Maps support-conditioned evidence to the explicit degree inference and possible defeaters.
Authentic assessment (Ajjawi et al.)	Makes tasks relevant to actual and evolving professional practice.	Whether authentic AI-supported success warrants independent or durable capability.	Pairs authentic work with bounded independence and transfer samples where the claim requires them.



Programmatic assessment (van der Vleuten et al.)	Aggregates longitudinal evidence for feedback and defensible decisions.	How to attribute performance across AI assistance conditions.	Prespecifies AI-support contrasts and an outcome-specific evidence ledger.
AI Assessment Scale (Perkins et al.)	Specifies and communicates permitted degrees of AI involvement.	Whether performances under permitted use justify the claimed competence.	Uses disclosed AI conditions as inputs to an inference audit, not a misconduct classifier.
AI literacy and evaluative judgment (Bearman et al.)	Defines quality of tool evaluation, questioning and oversight.	How these capabilities coexist with nondelegable domain skills and future transfer.	Evaluates independent and augmented capabilities separately, then adjudicates their credential significance.

3. METHODOLOGY: ANALYTICAL DESIGN-SCIENCE DEVELOPMENT

This is a conceptual and analytical design-science study, not an intervention trial, interview study, systematic review, or statistical meta-analysis. The method follows design-science principles for constructing a useful artifact against a specified problem (Peppers et al., 2007; Gregor & Hevner, 2013). Its evidence base is a purposive synthesis of established work on validity, retrieval and transfer, AI-enhanced teaching, educational assessment, and human-AI judgment. Sources were selected for direct relevance to the inferential problem and for complementary theoretical mechanisms, rather than according to a preregistered exhaustive search protocol. Thus, the paper makes no claims about evidence coverage or effect-size aggregation.

Development consisted of six logical operations. First, the problem was defined as potential separation of observed performance and credential-inferred competence. Second, required design functions were specified: construct alignment, support-condition contrast, transfer sampling, AI accountability, fairness, and low implementation burden. Third, alternative configurations were compared conceptually, including an AI ban, disclosure-only assessment, oral defense alone, and a multimode audit. Fourth, the CCA artifact was specified as a set of tasks, evidence records, decision rules, and program feedback loops. Fifth, three explicitly hypothetical cases were used to probe foreseeable false positives, false negatives, and governance failures. Sixth, observable propositions and a pilot study capable of refuting the design assumptions were articulated. The work therefore demonstrates analytic coherence and feasibility considerations, not empirical effectiveness.

A transparent design assumption underpins the audit: no short assessment can identify an individual's true competence without error. The aim is to improve the warrant for a bounded credential claim by combining multiple relevant observations, while making residual uncertainty and possible sources of bias visible. Each university must localize the artifact to disciplinary standards and regulatory requirements; no universal percentage weighting is asserted.

Traceability is treated as a design criterion: every proposed inference must name its target outcome, the performance mode that samples it, the observable rubric criterion, the exception or rival explanation that could defeat it, and the decision for which the evidence is used. This is a design specification, not evidence that the resulting decisions are valid. Initial pilots should compare CCA with an appropriately strong existing assessment system, rather than with an artificially weak single essay.

The credential mirage can be stated operationally as an inference mismatch: the institution or its stakeholders claim a specified level of graduate capability that is not warranted by the available, properly interpreted assessment evidence. It does not follow from the size of an AI-related score gap alone. An audit finding can establish insufficient warrant for a claim, but cannot by itself prove that AI caused the gap, diagnose the student's underlying ability, or estimate the prevalence of the phenomenon.

Table 2 defines the constructs that delimit the credential claim and identifies observable indicators for each.



Table 2. Core constructs and their observable indicators

Construct	Working definition	Possible evidence
Artifact quality	Quality of the finished work irrespective of how inputs were supplied.	Accuracy, relevance, technical sophistication and clarity of the submitted output.
Bounded independent competence	Ability to carry out indispensable reasoning with predeclared non-generative resources and appropriate accommodations.	Individual explanation, diagnosis, justified choice, error correction.
Accountable augmentation	Ability to select, challenge, document and responsibly use AI assistance.	Tool justification, error detection, provenance explanation, calibrated reliance.
Delayed transfer	Capacity to redeploy learning to a changed task after a meaningful interval.	Novel problem-solving with altered assumptions and retained conceptual rationale.
Credential validity	Strength of the evidence warranting an award-level capability claim.	Consistent multimode evidence, documented exceptions and fair decision rules.

4. THE CREDENTIAL MIRAGE AND ITS MECHANISMS

The credential mirage occurs when stakeholders interpret observable educational achievement more broadly than the evidence allows. It is a conditional risk, not a measured prevalence rate. Four distinguishable mechanisms can produce this risk. Performance substitution occurs when an AI system supplies the essential cognitive work represented by an assessment. Attribution opacity arises when faculty cannot determine which competencies belong to the student or the tool. Transfer discontinuity appears when success on a familiar AI-supported task does not generalize to a genuinely changed task. Credential amplification occurs when the institution translates weak task-level evidence into strong graduate-level claims, especially in externally marketed promises about critical thinking, independent expertise, or professional readiness.

These mechanisms interact but are not equivalent. A student may independently understand a solution, use AI to improve prose, and still possess the underlying competence. Another may skillfully select AI tools but lack a core domain skill required for safe practice. A third may fail an oral defense because of language anxiety or disability despite demonstrating competence under accessible formats. Accordingly, the audit should test intended constructs and competing explanations rather than treat a change in performance as conclusive proof of dependency.

Figure 1 distinguishes the observable artifact, the scoring event, and the external credential inference. The audit inserts evidence at the inference boundary: bounded independence, accountable augmentation, and delayed transfer are complementary witnesses to competence. The purpose is not to produce three redundant grades but to locate where the evidentiary chain is strong, ambiguous, or broken.

The three modes have deliberately different evidential jobs. Mode A supports a restricted inference about indispensable reasoning under specified non-generative supports. Mode B supports the inference that the learner can supervise, verify and take responsibility for AI-assisted work. Mode C supports a claim about retained adaptation after task and time changes. Convergence strengthens the claim only when tasks are aligned and accessible; disagreement identifies which specific inference needs further evidence rather than authorizing a generic dependence label. Thus, the modes are neither interchangeable nor automatically given equal weight.

Figure 1 | Credential validity is an inference, not an artifact score

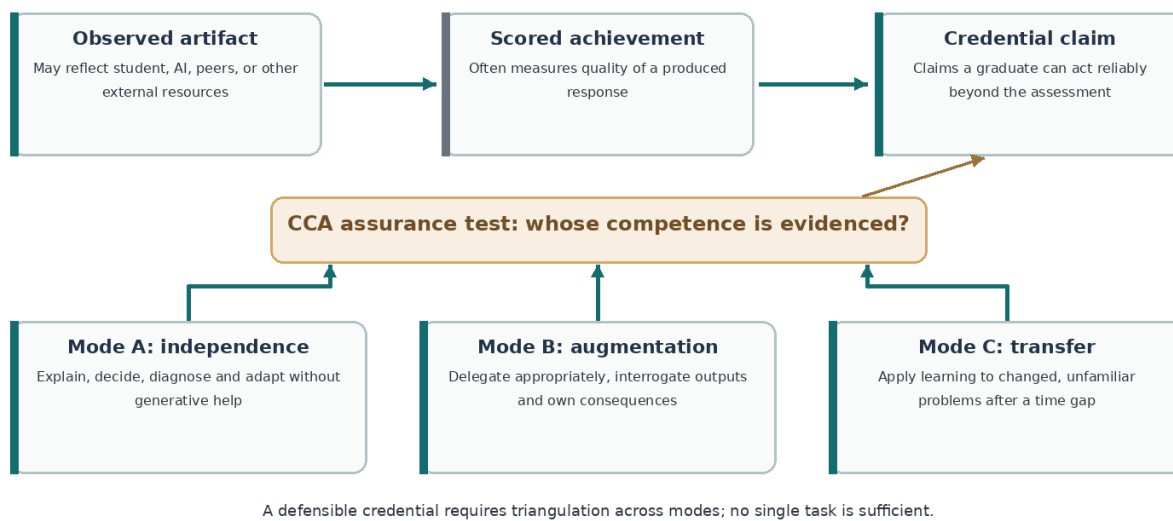


Figure 1. The credential mirage as an inference and triangulation problem.

Conceptual illustration developed for this study; not an empirical causal model or measured pathway.

5. COUNTERFACTUAL COMPETENCE AUDIT: DESIGN AND OPERATION

5.1. Six design principles

Principle 1, claim-first assessment: specify what the degree promises before selecting formats. Some outcomes, such as independent clinical triage or core mathematical diagnosis, may require bounded performance without generative output. Other outcomes, such as AI-supported market intelligence, require supervised tool use. Principle 2, explicit assistance contracts: state which tools, prompts, human collaboration, accommodations, and documentary aids are permitted in each mode. Disclosure should inform interpretation rather than be treated as a substitute for evidence (Perkins et al., 2024).

Principle 3, contrast without punishment: use assistance-condition variation as a diagnostic, not as an automatic misconduct trigger. A lower score in a no-AI task may reflect anxiety, lack of familiarity, language barriers, or poor construct alignment. Principle 4, delayed and changed-context sampling: include at least one application task that requires transformation rather than reproduction. Principle 5, accountability through judgment: request explanations of trust calibration, verification steps, rejected suggestions, sources, and ethical consequences. Principle 6, program-level learning: aggregate patterns to redesign curriculum and task portfolios rather than continuously surveil individuals.

5.2. The three-mode evidence portfolio

Mode A is bounded independence. It samples indispensable skills using resources specified in advance, such as an approved reference sheet, calculator, assistive technology, or domain handbook, while limiting generative completion. The precise boundary should reflect actual professional safety and curricular claims; a blind closed-book examination is not a universal default. Mode B is accountable augmentation. AI access is permitted or required, but students must show tool selection, quality checks, the rationale behind final decisions, documented uncertainty, and responsibility for consequential errors. Mode C is delayed transfer. After a stated interval, students face an analogous but materially altered case; this may occur with AI conditions chosen to match the exact transfer claim being tested. A delayed task tests durability and adaptation, not solely memory.

These modes can be administered as distinct tasks, rotated stations, or a small representative portfolio. For



example, a management capstone could require an independently reasoned recommendation, an AI-supported scenario analysis with an error log, and a subsequent defense of a changed market-entry dilemma. A computing module could contrast unaided debugging principles, AI-assisted code review, and application to an unfamiliar defect class. A nursing program would define safety-critical judgments under the relevant licensing standards. In each case, the institution should avoid creating duplicate high-stakes examinations where existing tasks can be adapted.

Table 3 shows how the three evidence modes may be translated into discipline-sensitive task designs.

Table 3. Illustrative multimode assessment designs by discipline

Program	Mode A: bounded independence	Mode B: accountable augmentation	Mode C: delayed transfer
MBA / management	Interpret a novel strategic trade-off and defend assumptions.	Use AI for scenario construction; document weak premises and rejected options.	Revisit the case after a competitor shock or regulatory change.
Software engineering	Trace a logic flaw and justify a fix under specified reference conditions.	Use an assistant to compare implementations; validate security and tests.	Debug a structurally different failure after an interval.
Health professions	Reason through a safety-relevant decision within permitted clinical resources.	Evaluate AI decision support with reference to professional standards.	Address an altered patient profile using a separately validated, accessible protocol.
Humanities / social science	Interpret an unfamiliar source and build a warranted argument.	Critique AI-generated synthesis, bias, missing perspectives and citations.	Reapply the conceptual lens to a new source or social context.

5.3. Traceability from principles to warrant and potential falsification

Principles become testable only when they determine observable features and different decisions. Table 4 provides the missing traceability layer: principles govern assessment design, modes supply evidence, rubric criteria describe its quality, decision rules protect the claim, and propositions evaluate whether this architecture works. 'P1–P5' refer to the prospective hypotheses specified in Section 7; they are not empirical results.

Table 4. Design-principle traceability across evidence, decision and proposition

Design principle	Evidence source / mode	Rubric indicator	Decision implication	Test link
D1. Claim-first	Outcome map; A/B/C chosen by claim.	Construct-relevant independent reasoning or supervision.	No inference broader than sampled outcome.	P1, P2
D2. Explicit assistance contract	Tool, source and access log; chiefly B.	Provenance, contribution, justified tool reliance.	Interpret only under declared support conditions.	P2, P4
D3. Contrast without punishment	Matched A/B tasks and access checks.	Mode-specific evidence, not raw score gap.	Discordance triggers inquiry, never automatic misconduct finding.	P1, P4
D4. Delayed changed-context sampling	Time-stamped novel C task.	Adaptation, retention, changed-constraint response.	Transfer claim requires evidence at matched difficulty.	P1, P3
D5. Accountability through judgment	B error-injection, rationale and checking record.	AI evaluation and decision ownership.	Augmented-competence claim needs verified oversight.	P2



D6. Program-level learning	Moderation, cohort trends, burden/trust logs.	Stable interpretations across assessors and groups.	High-stakes use withheld if calibration or equity fails.	P4, P5
----------------------------	---	---	--	--------

5.4. Evidence record, rubric and decision logic

The CCA produces a claim-to-evidence register for each selected program outcome. Each record contains the outcome statement; indispensable versus delegable capabilities; task specification and difficulty; permitted AI, reference, assistive and human supports; assessment date and transfer interval; rubric anchors; assessor identity and moderation; relevant accommodations; adverse conditions; and a concise rationale linking evidence to the proposed inference. Four qualitative anchors are used consistently: insufficient, emerging, defensible and strong. 'Insufficient' means the available observation does not support the proposed claim and may reflect missing or invalid evidence, not necessarily low ability. Anchors are illustrative, require local standard-setting, and must not be mistaken for validated universal cut scores.

Table 5 operationalizes qualitative evidence levels; the anchors are proposals requiring local calibration.

Table 5. Proposed scoring anchors for CCA evidence

Rubric dimension	Insufficient	Emerging	Defensible	Strong
Independent reasoning (A)	Critical reasoning not observable, or conditions invalidate judgment.	States a partial rationale but misses an essential assumption.	Explains a justified choice and responds to a plausible challenge.	Generalizes the reasoning and diagnoses multiple failure conditions.
AI evaluation (B)	Cannot assess central AI claims or consequential errors.	Checks obvious outputs but misses salient uncertainty or bias.	Verifies high-stakes claims and documents limitations and rejected options.	Calibrates reliance, independently tests alternatives and corrects subtle errors.
Delayed transfer (C)	No comparable changed-context performance is available.	Applies familiar steps with little adaptation to changed conditions.	Adapts principles to a materially different task after a stated delay.	Adapts across competing constraints and justifies counterfactual decisions.
Provenance and ownership (B)	Cannot establish meaningful learner contribution or responsibility.	Reports tool use incompletely or with weak attribution.	Names tools, own contribution, checking steps and decision ownership.	Maintains auditable rationale and proportionate escalation of uncertainty.

A key decision rule is noncompensation for safety-critical outcomes: exceptional AI-supported output cannot, by itself, offset a failure to demonstrate an essential independent safety competence. For other outcomes, a low independent score need not invalidate the qualification if the published construct explicitly concerns collaborative human-AI performance and the tool-supervision evidence is compelling. Divergent evidence triggers targeted re-assessment or learning support, not an automated sanction. Individual grades remain governed by published local assessment policy; the CCA initially operates as a quality-assurance overlay rather than a retroactive new grading system.

The audit should also define stopping conditions. If two appropriately sampled tasks are contaminated by undisclosed external assistance, or if access conditions differ in ways that cannot be resolved, the correct outcome is “insufficient interpretive evidence,” not a numerical estimate of diminished ability. Multiple

assessors and periodic moderation can help prevent the rhetoric of precision from exceeding the reliability of the observed evidence.

5.5. Claim-specific standard setting, weighting and conflict decisions

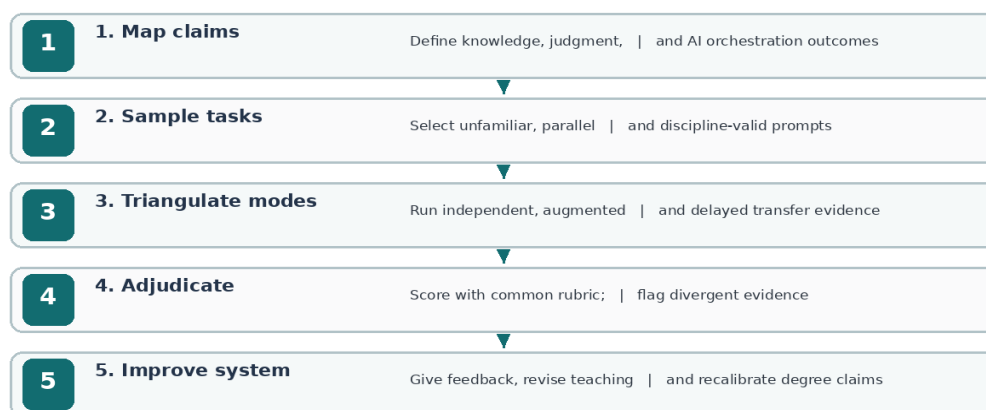
CCA does not prescribe a universal A/B/C weighting. Before testing, a disciplinary panel publishes an outcome-by-mode blueprint: the claim being certified, which competencies cannot be delegated, which require accountable AI use, the purpose and interval of delayed transfer, acceptable equivalent formats, evidence sufficiency, and the action attached to each possible result. A program that promises independent diagnostic reasoning must treat adequate A evidence as indispensable. A program promising responsible human-AI orchestration must give B substantial interpretive importance. When durable adaptation is central, C must provide independent supporting evidence. These are rules governing inferential relevance, not arbitrary equal weights.

Initial standard setting should involve representative faculty, assessment specialists, professional or employer advisors where appropriate, disability-access experts, and student input. Panels examine anchor examples and borderline performances, set provisional outcome-specific standards, and document the justification and acceptable uncertainty. Pilot data are then used to examine parallel-form comparability, rater agreement, response process evidence, differential performance across groups and temporal stability, followed by a recorded revision decision. No compensatory formula should override a prespecified safety-critical or regulated capability. This approach follows the emphasis on justified interpretation and defensible decisions in contemporary assessment standards (American Educational Research Association et al., 2014; Kane, 2013).

For each claim, the adjudication sequence is: (1) verify task alignment and accessible support conditions; (2) inspect A, B and C evidence for the outcome they were designed to sample; (3) mark any nondelegable capability as a separate hurdle; (4) reconcile discordance using a second task, assessor or construct-equivalent response format; and (5) issue an outcome-specific conclusion. Permissible audit conclusions are 'supported under specified conditions', 'supported for augmented practice but not for an unassisted claim', 'further evidence required', or 'not supported by the present evidence'. An individual award decision remains subject to the institution's prepublished regulations and right of appeal; a formative pilot must not retroactively change a student's qualification.

Missingness is not failure. If tool access differs, a transfer task is leaked, a disabled learner is assessed in an inaccessible format, or scoring agreement is inadequate, the observation is classified as invalid or indeterminate and resampled before consequential use. The burden of establishing inferential warrant rests with the institution, not with a student disproving presumed AI misuse.

Figure 2 | Counterfactual Competence Audit (CCA) workflow



Cycle repeats by program and cohort; aggregate results inform quality assurance, not student surveillance.

Figure 2. Operational cycle of the Counterfactual Competence Audit.



Proposed institutional workflow; steps are iterative, and local validation should precede high-stakes use.

6. EX ANTE STRESS TESTS AND ANALYTICAL RESULTS

The following are thought experiments, not observed student cases. They test whether the artifact yields plausible distinctions under contrasting conditions. The analytic criterion is whether the CCA identifies when evidence is insufficient and when legitimate AI-enabled capability can be recognized without overclaiming unaided mastery.

Case A: the excellent AI-mediated essay, weak independent interpretation

A student submits a persuasive policy essay, openly identifies substantial AI assistance, and earns a high quality score. When asked to explain an unfamiliar counterargument using an accessible short-response task, the student cannot identify the crucial assumption; a later transfer task shows similar difficulty. A disclosure-only system may correctly judge the essay but cannot substantiate a strong claim about independent disciplinary interpretation. The CCA would flag a gap between artifact quality and Mode A/C evidence, supply formative instruction, and qualify the program-level competence inference. It would not infer dishonesty from authorized AI usage.

Case B: modest independent fluency, exceptional accountable augmentation

An experienced working student writes slowly but makes sound independent decisions under reasonable accommodations. With AI, the student produces a substantially stronger analytic product, detects a fabricated source, chooses better evidence, and can explain why certain suggestions were rejected. A no-AI-only regime could undervalue this genuine workforce capability. The CCA recognizes strong Mode B evidence while preserving a narrow independent claim about the competencies actually demonstrated. It therefore avoids treating all AI assistance as substitution.

Case C: apparent performance divergence caused by inequitable conditions

Two learners show similar task understanding, but one faces accessibility barriers in an improvised oral defense while the other uses an accessible written explanation. A crude “AI versus no AI” score gap could mistakenly label the first learner dependent. Under the CCA, access conditions, comparability, and construct relevance are recorded before adjudication. The apparent divergence would remain uninterpretable until comparable accessible assessments were available. This is a central failure-avoidance function, not a peripheral accommodation.

These cases support an analytic result rather than empirical validation: a three-mode scheme can express distinctions that single-mode scoring or tool-disclosure rules cannot express cleanly. They do not prove that the scheme raises validity, predicts employment outcomes, or reduces inequity. Those claims require field evaluation and independent replication.

Case D: a worked MBA market-entry capstone

Consider a hypothetical MBA program whose advertised outcome is: 'Graduates can independently identify strategic risks, responsibly use AI for option analysis, and adapt recommendations after market disruption.' A learner submits a polished AI-assisted international market-entry report. The task includes a deliberately misleading market statistic generated by the AI. The learner spots the problem and corrects it, but a separate accessible independent task reveals an incomplete explanation of a regulatory entry constraint. Four weeks later, the learner adapts much of the strategy after an unexpected tariff, yet overlooks the same constraint. The record below is fictional and uses illustrative qualitative anchors, not real student data, institutional cutoffs or a proposed automatic grading algorithm.

Table 6. Hypothetical CCA evidence ledger and decision path for an MBA capstone

Evidence / claim	Illustrative observation	Anchor	Warranted inference	Adjudication
Artifact score	Persuasive AI-supported market-entry report.	Strong output (not a CCA anchor)	Quality of the artifact, not independently proven competence.	Retain ordinary assignment mark under existing rules.
Mode A: nondelegable risk diagnosis	Explains commercial trade-offs but misses the regulatory constraint.	Emerging	Insufficient present warrant for the specified independent risk outcome.	Provide feedback and a comparable reassessment opportunity.
Mode B: AI oversight	Finds false market statistic, checks sources, rejects spurious advice.	Strong	Evidence for accountable AI-augmented strategic analysis.	Support the specifically augmented claim.
Mode C: transfer after 4 weeks	Revises strategic options after tariff change; repeats regulation omission.	Emerging	Some adaptation but incomplete support for durable regulatory reasoning.	Use a parallel transfer task after targeted learning.
Program-level interpretation	A/C conflict with high artifact and B evidence.	Mixed	No defensible inference of fully independent strategic risk diagnosis yet.	Record further evidence required; no misconduct conclusion.

This example shows why a single overall weighted total could be misleading: a strong B score should not compensate for an indispensable A capability that remains unsupported. It also avoids the opposite mistake of treating weak A performance as proof that AI-dependent students lack valuable augmented competence. After equitable resampling, evidence might support the independent outcome, or the institution might need to narrow its advertised claim. Neither outcome is predetermined by the audit. The proposed record is falsifiable because a prospective study can test whether these decisions better predict later external performance than existing program assessments.

Table 7 summarizes foreseeable validity threats and safeguards that require further evaluation.

Table 7. Threat model and proposed safeguards

Threat to validity	Potential distortion	Preventive control and remaining uncertainty
AI-created apparent mastery	Artifact quality conflated with learner reasoning.	Match independent explanation to explicit claim; hidden external help remains possible.
Blanket AI restriction	Workplace-relevant oversight is not sampled.	Assess accountable augmentation under a transparent assistance contract.
Disability and assistive-technology barriers	An oral, speeded or inaccessible task measures access rather than intended capability.	Offer preplanned construct-equivalent formats, accessible tools and individual adjustments; study comparability.
Language and cultural differences	Fluency, accent or culturally narrow scenarios confound disciplinary reasoning.	Minimize construct-irrelevant language load and moderate translated or alternative responses.
Unequal finances, devices or AI access	Paid model access and home resources affect B mode performance.	Supply a common institution-controlled baseline, accessibility, practice time and transparent tool conditions.



Task leakage and coaching	Rehearsal appears as transfer.	Use parallel task families, independent moderation and novel perturbations; monitor difficulty drift.
Model updates and interface changes	Different AI systems or dates change task challenge.	Version and record tool conditions, re-anchor rubrics and repeat comparability checks.
Excessive monitoring and inference	Privacy violations and student distrust.	Prefer sampled evidence and minimal records; avoid device surveillance and undisclosed profiling.
Assessor inconsistency	Judgments differ by reviewer or discipline.	Anchor moderation, assessor training, duplicate-scoring samples and adjudication trails.
Overstated credential marketing	Claims exceed observed constructs.	Publish narrow outcomes, support conditions and explicit uncertainties.

7. RESEARCH PROPOSITIONS AND FALSIFICATION AGENDA

The following conditional propositions are distinct empirical claims. P1 (incremental criterion validity) predicts that CCA evidence will improve out-of-sample prediction of a later novel, discipline-relevant independent task beyond prior marks, baseline skills and declared AI exposure. P2 (augmented oversight) predicts that Mode B evidence will predict detection of seeded AI errors and appropriately calibrated reliance in a later AI-assisted task beyond AIAS permission levels and overall artifact quality. P3 (durable transfer) predicts that Mode C will provide additional discrimination for subsequent altered-context performance beyond immediate-task marks, after controlling baseline competence and elapsed time.

P4 (equity with construct preservation) predicts that accessible, construct-equivalent administration will reduce format-attributable subgroup gaps without sacrificing criterion-relevant construct coverage or future-task prediction. P5 (governance proportionality) predicts that a sampled CCA program audit will produce actionable curricular signals at lower staff time, intrusive-data collection and student trust cost than a prespecified intensive-monitoring comparator. None is tautologically guaranteed: evidence of no useful incremental prediction, diminished fairness, unstable measurement, or excessive cost would count against the respective proposition.

A prospective multi-program pilot should preregister each claim, task family, anticipated direction, smallest educationally meaningful improvement, comparator and analytic plan. Where feasible, randomly assign assessment order and parallel-form versions, while preserving access arrangements and avoiding denial of course credit. Record prior achievement, baseline AI literacy, language and accessibility conditions, socioeconomic access proxies only with appropriate consent, and available AI tools. Blind external assessors to CCA classifications. Estimate agreement, generalizability across forms, predictive calibration, false-positive and false-negative claims, subgroup differential functioning, trust, completion time and staff burden. Report uncertainty intervals, failed hypotheses and contextual limitations. A university must stop short of individual high-stakes deployment unless the local design meets reliability, fairness and appeals standards.

Table 8 connects the framework's empirical claims with prospective tests and disconfirming observations.

Table 8. Proposed prospective validation plan

Hypothesis / comparison	Measured endpoint	Comparator and disconfirmation
P1: CCA adds criterion validity to grades.	External changed-task score; out-of-sample predictive gain.	Standard program assessment plus baseline AI exposure; no reliable predictive gain or poor cross-cohort replication.



P2: B measures accountable augmentation.	Seeded-error detection, justified reliance and downstream decisions.	Artifact quality and AIAS condition; no incremental prediction of responsible AI behavior.
P3: C samples durable transfer.	Altered delayed task performance after specified interval.	Immediate marks and baseline competence; negligible incremental discrimination after cost adjustment.
P4: accessible modes improve equity.	Format-specific gaps, task equivalence, criterion prediction by subgroup.	Existing format; gap persists or increases, or criterion coverage deteriorates.
P5: proportionate program audit is actionable.	Curricular changes, assessor hours, data collected and trust indicators.	Prespecified intensive-monitoring or conventional QA protocol; no useful improvement or excess burden/harm.

8. DISCUSSION

8.1. The meaning of an AI-resilient university degree

The objective is not to create an AI-proof examination. It is to make a degree claim resilient to changes in the technological source of an answer. A strong credential would indicate which capabilities graduates can demonstrate independently under specified reasonable supports, which they can perform through accountable human-AI collaboration, and which they can transfer across contexts. This approach treats uncertainty about AI contribution as a validity-design question rather than a default accusation against students. It also clarifies that an advanced degree cannot be reduced to fluency in prompt construction or mastery of fixed encyclopedic content.

Institutional communication should follow the evidence. A marketing claim such as “graduates are independent strategic thinkers” is not warranted solely by AI-assisted essays or polished capstones. A claim such as “graduates can critically deploy AI-enabled analysis and defend the resulting recommendations” requires actual evidence of verification and responsibility. This distinction aligns with the human-centered orientation of UNESCO guidance (Miao & Holmes, 2023) and competency frameworks emphasizing critical judgment and ethical practice (Miao et al., 2024). Employers would benefit from clearer descriptions of assessed capability, but any employer inference still requires attention to contextual differences between campus and workplace tasks.

8.2. Governance and implementation feasibility

A realistic implementation would audit two or three high-value program outcomes per cycle rather than reconstruct every assignment. Program teams could map existing capstone activities onto the three modes, use a shared assessor rubric, and sample students in quality-assurance studies with appropriate notice and consent processes. Decision owners should include teaching faculty, assessment specialists, accessibility services, and student representatives. Governance needs explicit restrictions on collecting prompts, chat transcripts, and personal device data; much of the needed evidence can be obtained through work products, brief reasoning records, and blinded task performance without intrusive monitoring.

The protocol should also distinguish two different uses of scores. For formative quality assurance, aggregated evidence may support curriculum improvement even when it is not reliable enough to penalize individuals. For high-stakes individual credential decisions, additional psychometric evidence, appeals, moderation, accessibility review, and prospectively published rules are essential. Institutions with limited resources should begin with a pilot and avoid introducing burdensome new defenses that disproportionately harm part-time learners, multilingual learners, or students relying on assistive technology. If these safeguards cannot be met, CCA should remain developmental rather than summative.

Equity should be designed in before tasks are used. Disability adjustments may include screen-reader compati-



bility, flexible timing, alternative communication modes, approved assistive AI where construct relevant, and comparable reference resources. Multilingual students should not be penalized for accent, idiom or speed when the outcome concerns strategic reasoning rather than oral English proficiency. Socioeconomic differences require baseline access to institution-provided AI tools, suitable devices, reliable connectivity and practice opportunities. For every accommodation, the university records what construct is unchanged, which support differs, and what evidence is needed to establish equivalence. Universal Design for Learning provides practical principles for reducing avoidable barriers (CAST, 2024); inclusion does not mean waiving genuine occupational safety standards.

Cross-disciplinary comparisons should concern the validity of common procedures, not identical raw scores. A nursing safety hurdle, an MBA regulatory-risk argument and a computing security diagnosis differ in stakes, resources and essential autonomous capabilities. Comparable audit quality can instead be assessed through whether programs publish a claim map, anchor assessments, demonstrate appropriate rater agreement, document accessibility, disclose AI conditions and follow consistent appeal rules. Institutional benchmarking of composite 'CCA scores' would be premature before invariance and criterion studies.

8.3. Relation to prior scholarship and unique contribution

CCA should be evaluated as a synthesis with one proposed operational increment, not as a wholly new theory of validity. Kane (2013) and Messick (1995) provide the inference and consequence logic; authentic assessment emphasizes future-relevant tasks (Ajjawi et al., 2024); programmatic assessment supplies evidence aggregation for decisions (van der Vleuten et al., 2012); and AIAS specifies legitimate assistance conditions (Perkins et al., 2024, 2025). CCA places contrastively sampled A/B/C evidence within those established traditions and makes the relation among support conditions, noncompensatory claims, accessible re-assessment and degree language explicit. Its originality claim would be weakened if existing well-implemented assessment systems already achieve equivalent calibration and prediction at lower cost. This is an empirical question, not an assumed advantage. Tan (2026a) concerns verification capacity in learning assurance, whereas the present proposal specifies and tests claim-level credential adjudication.

8.4. Limitations and boundary conditions

This conceptual design-science work does not establish the frequency of the credential mirage, the effectiveness of CCA, the correct A/B/C weights, or generalizable thresholds. The literature synthesis is purposive, not systematic. The illustrative MBA case, rubric anchors and decisions are hypotheses about a workable design, not validation results. An adapted task can remain confounded by motivation, intervening learning, disability access, unequal AI resources and differences in task difficulty; even well-calibrated modes may lack adequate incremental value. Important legal and professional rules vary by discipline. The protocol may intensify workload or surveillance if implemented poorly. High-stakes deployment therefore requires local prospective standard setting, measurement evaluation, stakeholder governance, fair appeals and external replication; until then CCA is a formative assurance proposal.

9. CONCLUSION

When AI helps students submit excellent work, the achievement may be real, but the inference from that artifact to a graduate's durable competence is not automatic. The credential mirage names the risk of making more expansive educational claims than available evidence supports. The Counterfactual Competence Audit offers a feasible conceptual response: define the claim, sample bounded independent and accountable AI-augmented competence, test delayed transfer, adjudicate uncertainty fairly, and feed results back into program quality assurance. The proposal neither idealizes unaided work nor assumes that AI support necessarily weakens learning. It asks universities to certify the capabilities they actually examine. Its ultimate value will depend on whether prospective studies demonstrate better credential validity, equitable practice, and acceptable implementation costs.

DECLARATIONS

Research design and data availability. This manuscript is a conceptual design-science contribution. It reports no newly collected participant data, field trials, or experimental estimates. The figures, tables, rubrics and illustrative cases are original proposals rather than observations.

REFERENCES

1. Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (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>
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. <https://www.testingstandards.net/>
3. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
4. Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, 49(6), 893–905. <https://doi.org/10.1080/02602938.2024.2335321>
5. Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, 31(4), 399–413. <https://doi.org/10.1080/02602930600679050>
6. Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(5), 1118–1133. <https://doi.org/10.1037/a0019902>
7. CAST. (2024). *Universal design for learning guidelines version 3.0*. <https://udlguidelines.cast.org/>
8. Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
9. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
10. Kestin, G., Miller, K., Kiales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, Article 17458. <https://doi.org/10.1038/s41598-025-97652-6>
11. 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>
12. Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
13. Miao, F., Shiohira, K., & Lao, N. (2024). *AI competency framework for students*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000391105>
14. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>
15. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
16. 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). <https://doi.org/10.53761/q3azde36>
17. Perkins, M., Roe, J., & Furze, L. (2025). Reimagining the Artificial Intelligence Assessment Scale: A refined framework for educational assessment. *Journal of University Teaching & Learning Practice*, 22(7). <https://doi.org/10.53761/rrm4y757>
18. 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>



19. Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255.
<https://doi.org/10.1111/j.1467-9280.2006.01693.x>
20. Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778.
<https://doi.org/10.1126/science.1207745>
21. Tan, K. H. (2025a). The AI-augmented MBA paradigm: Reconceptualizing business education relevance in the era of artificial intelligence. ResearchGate.
<https://doi.org/10.13140/RG.2.2.24863.06566>
22. Tan, K. H. (2025b). The transferable skills premium: Redefining doctoral degree value in the knowledge economy. ResearchGate. <https://doi.org/10.5281/zenodo.17336334>
23. Tan, K. H. (2025c). The exploration of ethical considerations in pedagogy. ResearchGate.
<https://doi.org/10.13140/RG.2.2.30940.71048>
24. Tan, K. H. (2026a). The learning assurance frontier: A capacity-constrained theory of generative AI, knowledge verification, and independent learning transfer. *Journal of Advanced Multidisciplinary Studies*, 16–31. <https://doi.org/10.68050/JAMS.2026.762>
25. Tan, K. H. (2026b). Improving critical thinking skills among adult learners through structured reflective dialogue: An action research study in a university social sciences course in Singapore. *International Journal of Research Publication and Reviews*, 7(6), 5322–5329.
<https://doi.org/10.55248/gengpi.07.0626.17a03>
26. van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205–214. <https://doi.org/10.3109/0142159X.2012.652239>