Stage 3 REVIEW —Peer Review Report

Paper: "From Snapshots to Trajectories: How Agentic AI Will Redefine Student Learning Outcomes and Transform Student Success Measurement —Implications for Taiwan's Next Cycle of Institutional Accreditation"

Date: 2026-03-07

Review Round: 1

Reviewer 1 (R1): Methodology Expert

Overall Assessment

This paper attempts an ambitious theoretical synthesis, proposing the ADAPT framework as an original conceptual contribution for understanding how agentic AI might transform student learning outcome measurement in Taiwan's higher education quality assurance system. The four-layer analytical framework —conceptual taxonomy, Kuhnian paradigm shift analysis, Bardach's eightfold path, and principlist ethical evaluation —is intellectually sophisticated, and the attempt to integrate these layers through a single unifying framework (ADAPT) is commendable. The paper demonstrates wide reading across assessment theory, AI scholarship, quality assurance literature, and educational policy.

However, there are significant methodological concerns that must be addressed before this paper meets the standard of a Q1 journal. The most fundamental issue is the tension between the paper's ambitious theoretical claims and the thinness of its evidentiary base. A theoretical paper is not exempt from evidentiary standards —it is held to different ones, and this paper does not always meet them. The Kuhnian framework, while productively applied, is stretched in ways that require more careful justification. The ADAPT framework, while elegantly constructed, remains largely stipulative rather than derived, and its internal logic needs tightening. The paper's five research questions are well-formulated but unevenly addressed, with RQ3 (the seven dimensions of paradigm shift) receiving disproportionate attention relative to RQ1 (definitional) and RQ5 (ethical governance).

Strengths

1. **Rigorous multi-framework architecture.** The integration of Kuhn, Bardach, principlist ethics, and conceptual taxonomy construction into a coherent analytical structure is genuinely original and well-executed. Each framework is applied to the domain for which it is best suited —Kuhn for the paradigm shift argument, Bardach for policy analysis, Beauchamp & Childress for ethics. This demonstrates methodological maturity and avoids the common trap of forcing a single framework to do work it cannot support.
2. **The seven dimensions of paradigm shift (Section 4.3) are well-constructed.** Each dimension is clearly defined, the contrast between current and agentic paradigms is sharply drawn, and the "mechanism of change"column in Table 1 (Section 4.3) usefully specifies how agentic AI capabilities enable each shift. The parallel structure of the analysis —each dimension discussed with benefits and risks — demonstrates balanced argumentation.
3. **Section 7.2 ("What This Paper Does NOT Claim") is exemplary.** Explicitly delineating the paper's epistemic boundaries —no empirical claims, no technology determinism, no claim that the current

paradigm is worthless, no claim that AI should replace human judgment —is a mark of intellectual honesty that strengthens rather than weakens the paper's authority.

4. **The Wei-Lin scenario (Section 4.4) is an effective methodological device.** Grounding abstract theoretical claims in a concrete, institutionally specific narrative helps the reader evaluate whether the paper's theoretical commitments translate into plausible practice. The scenario is appropriately qualified as "deliberately realistic"rather than speculative.

5. **The paper's reflexive acknowledgment of limitations (Section 7.3) is thorough.** Five specific limitations are identified, each with genuine analytical bite. This is not pro forma self-criticism but substantive methodological reflection.

Weaknesses

1. **The Kuhnian framework is applied without adequate justification of its applicability to educational assessment.** Kuhn's theory was developed for the natural sciences; its extension to social practices like quality assurance requires explicit argument about why the concepts of "normal science,""anomaly,""crisis,"and "incommensurability"apply in this domain. The paper acknowledges that "scholars have productively applied [Kuhn's framework] to education (Shepard, 2000)"(Section 4.1) but does not engage with the substantial literature critiquing such extensions. Kuhn himself was skeptical of applying his framework outside the natural sciences. Is Taiwan's QA system really a "paradigm" in the Kuhnian sense —a shared framework of exemplars, methods, and ontological commitments? Or is it better characterized as an administrative regime subject to policy reform? The paper needs to address this question head-on, or it risks the charge that the Kuhnian framing is metaphorical rather than analytical.

2. **The ADAPT framework is stipulative rather than derived.** The paper announces the five components of ADAPT —Agency Architecture, Diagnostic Mapping, Assessment Reconception, Policy Pathways, Trust & Ethics Safeguards —but does not adequately explain why these five and not others. What is the derivation logic? Is it empirical (derived from case analysis), theoretical (derived from first principles), or pragmatic (derived from the research questions)? The paper suggests the latter (Section 4.2: "integrates the paper's five research questions into a unified analytical structure"), but this makes ADAPT a re-labeling of the research questions rather than an independent analytical contribution. A stronger justification would demonstrate that these five components are necessary and sufficient for analyzing the paradigm shift —or at minimum, explain the selection criteria.

3. **The evidence base for agentic AI capabilities is thin and selectively cited.** Section 2 relies heavily on Tao et al. (2026), Masterman et al. (2024), El-Banna et al. (2025), and Gartner (2025) —sources that are either industry reports, preprints, or very recent publications without established citation impact. The Kestin et al. (2025) study from Harvard is cited repeatedly as evidence of AI tutoring effectiveness, but the paper itself acknowledges its limited generalizability (Section 6.1.2). For a theoretical paper, the quality and diversity of the evidence base for the technological claims is crucial, and here it is insufficient. The paper needs either more robust evidence or more explicit hedging about the maturity of the technology.

4. **The relationship between Bardach's eightfold path and the actual policy analysis in Section 5 is unclear.** The paper claims Bardach as a methodological pillar but does not systematically walk through the eightfold path's steps (define the problem, assemble evidence, construct alternatives, select criteria, project outcomes, confront trade-offs, decide, tell your story). Section 5 does some of this implicitly, but the framework is not made visible in the analysis. If Bardach is a core methodological commitment,

it should be rigorously applied; if it is merely an inspiration, the paper should say so.

5. **The paper uses two different tables both labeled "Table 1."** Section 3.2 contains "Table 1: Summary of Student Learning Outcome Assessment Instruments"and Section 4.3 contains "Table 1: Seven Dimensions of the Assessment Paradigm Shift."This is a formatting error that should be corrected, but it also signals a broader issue with the paper's internal organization and editorial polish.

Specific Comments

- **Abstract (lines 6-20):** The abstract is long (approximately 450 words) and tries to preview every element of the paper. For a Q1 journal, a tighter abstract of 250-300 words would be more effective. The current version reads like an executive summary rather than an abstract.
- **Section 2 (Agentic AI taxonomy):** The four-level taxonomy (reactive, autonomous-single, adaptive-multi, ecosystem) is useful but would benefit from clearer boundary conditions. When does a system move from Level 2 to Level 3? What observable criteria distinguish levels? Without operational definitions, the taxonomy risks being unfalsifiable.
- **Section 4.1 ("Incommensurability"subsection):** The claim that "a graduation rate and a learning trajectory are not different measures of the same thing"(Section 4.1) is provocative but under-argued. One could reasonably argue they are different measures of the same underlying construct (educational quality) at different levels of granularity. The incommensurability claim needs more philosophical work.
- **Section 5.3 (Phased approach):** The three phases (2026-2028, 2028-2030, 2030+) are plausible but the transition criteria between phases are vague. What specific evaluation results from Phase 1 would trigger Phase 2? What would constitute evidence that "AI-augmented assessment can produce valid, equitable, and pedagogically valuable learning evidence"? Without operationalized transition criteria, the phased approach risks becoming a rhetorical device rather than a genuine policy recommendation.
- **Section 6.3 (Agentic-specific risks):** This is one of the paper's strongest analytical contributions. The distinction between general AI risks and agentic-specific risks (autonomy amplification, iterative reasoning unpredictability, persistent memory bias compounding, multi-agent coordination failures, goal misalignment) is well-drawn and genuinely advances the ethical discourse beyond generic AI ethics.

Questions for Authors

1. How do you respond to the critique that the Kuhnian framework is metaphorical rather than analytical in this context? What makes Taiwan's QA system a "paradigm"rather than an administrative regime?
2. What is the derivation logic for the five components of the ADAPT framework? Can you demonstrate that these five are necessary and sufficient, or at minimum explain the selection criteria?
3. How would you operationalize the transition criteria between the three implementation phases? What specific metrics or evaluation outcomes would trigger progression from Phase 1 to Phase 2?
4. You cite the Kestin et al. (2025) Harvard study as evidence of AI tutoring effectiveness. How do you address the generalizability concerns, given that the study was conducted in a specific, well-resourced context with a narrow outcome measure?
5. The paper proposes both formative and summative functions within an integrated agentic system (Section 4.3.6) but acknowledges the tension between them. How would you architecturally separate these

functions? Is this a design problem or a conceptual problem?

Recommendation: Major Revision

The paper has genuine intellectual merit and makes an original contribution through the ADAPT framework and the seven-dimension paradigm shift analysis. However, the methodological foundations —particularly the Kuhnian framework application, the ADAPT framework derivation, and the evidence base for technological claims —require substantial strengthening. The relationship between Bardach's framework and the actual policy analysis needs to be made explicit. These are addressable issues, and I am optimistic that a thorough revision could produce a publishable paper.

Reviewer 2 (R2): Domain Expert

Overall Assessment

As a researcher with extensive experience in Taiwan's higher education quality assurance system and familiarity with HEEACT's operational realities, I find this paper to be an impressively informed and largely accurate analysis of the current QA landscape, coupled with a forward-looking policy vision that is both ambitious and grounded. The paper demonstrates genuine understanding of HEEACT's accreditation structure, the regulatory relationship between MOE and HEEACT, the institutional diversity of Taiwan's HE sector, and the cultural dynamics that shape quality assurance practice. The six structural limitations identified in Section 3 are accurately diagnosed, and the policy scenarios in Section 5 are realistic representations of the options available to Taiwan's QA decision-makers.

That said, the paper has several domain-specific weaknesses that a specialist audience would identify. The characterization of faculty resistance is somewhat stereotyped. The treatment of the demographic crisis, while factually accurate, underestimates the complexity of its interaction with quality assurance. And some of the specific policy recommendations, while intellectually coherent, face practical obstacles that the paper does not adequately address. These are not fatal flaws, but they require attention to ensure the paper's credibility with its target readership —the very policy-makers and QA practitioners who would need to act on its recommendations.

Strengths

1. **Accurate and nuanced characterization of Taiwan's QA architecture (Section 3).** The paper correctly identifies the six structural limitations and situates them within the specific institutional context of HEEACT's three-cycle history. The description of the SAR process, the on-site visit protocol, the role of self-accrediting institutions, and the relationship between MOE and HEEACT is accurate and demonstrates genuine domain expertise. The paper avoids the common error of treating HEEACT's framework as monolithic; it recognizes the evolution across three cycles and the increasing emphasis on institutional self-improvement.
2. **The three policy scenarios (Section 5.1) are realistic and well-differentiated.** Scenario A (conservative integration) accurately captures the default position; Scenario B (framework evolution) aligns with the direction many HEEACT insiders have privately discussed; Scenario C (paradigm replacement) is correctly identified as aspirational but impractical in the near term. The comparative analysis matrix (Table 5.1) is a genuinely useful policy tool.

3. **The South Korean cautionary tale (Section 5.2) is well-chosen and well-analyzed.** This section demonstrates the kind of comparative policy awareness that strengthens the paper's credibility. The three lessons drawn —technological readiness vs. stakeholder readiness, equity as precondition, and genuine piloting —are directly applicable to Taiwan's context.

4. **The fourth-cycle recommendations (Section 5.4) are specific and actionable.** Table 5.2 (proposed Core Indicator revisions) demonstrates the level of specificity needed to be taken seriously by policy audiences. The conditional phrasing ("where deployed,""the institution may demonstrate") is politically astute, allowing for voluntary adoption without mandating technological commitment.

5. **The paper correctly identifies self-accrediting institutions as natural early adopters (Section 5.4.4).** This is a strategically important observation. Self-accrediting institutions (currently approximately 34 institutions, including all national research universities) have the regulatory flexibility, technical capacity, and institutional motivation to serve as pilot sites. The proposal for "AI-Ready Self-Accreditation Guidelines"is practical and feasible.

Weaknesses

1. **The treatment of the fourth-cycle timeline is imprecise.** The paper states that "the third cycle of institutional accreditation concludes in academic year 2025"(Section 5.4), but this requires clarification. The third cycle (112-114 academic years, i.e., 2023-2025) includes institutions evaluated in different years within the cycle. The design window for the fourth cycle is already underway —HEEACT typically begins framework design 2-3 years before the first cohort of institutions is evaluated. The paper should specify more precisely when the fourth-cycle design window opens and closes, as this affects the feasibility of its recommended timeline.

2. **Faculty resistance is characterized too monolithically.** Section 3.4 describes "widespread faculty skepticism toward OBE"but does not distinguish between different types of faculty resistance or between different institutional contexts. Research university faculty may resist for philosophical reasons (autonomy, academic freedom); teaching-focused university faculty may resist for practical reasons (workload, inadequate support). The paper's policy recommendations would benefit from acknowledging this heterogeneity and proposing differentiated strategies.

3. **The paper underestimates the political economy of HEEACT's decision-making.** HEEACT operates under MOE's policy direction but with significant operational autonomy. The recommendations in Section 5.4 assume a level of MOE-HEEACT coordination and policy coherence that does not always obtain in practice. Budget allocation for pilot programs, evaluator training, and infrastructure investment requires MOE commitment that goes beyond HEEACT's independent decision-making authority. The paper should explicitly address the governance dynamics between MOE, HEEACT, and institutions in the policy implementation pathway.

4. **The demographic crisis analysis (Section 3.4) does not adequately address its implications for institutional capacity to adopt AI.** The paper correctly identifies the demographic crisis as an "anomaly" straining the current paradigm, but does not sufficiently consider that the same crisis is depleting institutional resources —reducing tuition revenue, triggering staffing cuts, and consuming administrative attention with survival rather than innovation. Many of the institutions most affected by the demographic crisis are precisely those least capable of implementing agentic AI systems. The paper's policy recommendations need to account for this resource paradox.

5. **The UCAN assessment characterization is oversimplified.** Section 3.2 describes UCAN as "self-reported competency diagnostics"with "annual (optional for students)"frequency. While technically

accurate, this misses the evolution of UCAN over the past decade —its increasing integration with institutional IR systems, its growing emphasis on competency-based rather than satisfaction-based measurement, and MOE's ongoing efforts to improve its diagnostic validity. A more nuanced treatment would strengthen the paper's credibility with domestic readers familiar with UCAN's trajectory.

6. **The paper does not adequately address the language and cultural dimensions of AI deployment.** Taiwan's higher education operates primarily in Mandarin Chinese, with significant use of Taiwanese Hokkien in some institutional contexts. Agentic AI systems trained predominantly on English-language data may perform poorly in Chinese-language educational settings, particularly for nuanced assessment tasks like evaluating argumentative writing, critical thinking, or cultural competency. This is a significant practical barrier that the paper does not discuss.

Specific Comments

- **Section 3.1 (HEEACT framework description):** The paper correctly describes HEEACT's three-cycle evolution but could benefit from noting the shift from "pass/conditional pass/fail"in the second cycle to the improvement-oriented "accredited/conditionally accredited"approach in the third cycle. This evolution is relevant because it shows HEEACT is already moving in the direction the paper advocates (from accountability to improvement).
- **Section 5.4.1 (Core Indicator revisions):** The proposed revision to Core Indicator 3-2 is well-formulated, but the paper should note that Core Indicator descriptors are deliberately broad to allow evaluators interpretive flexibility. Overly specific descriptor language could constrain evaluator judgment rather than enhance it. The balance between specificity and flexibility is a real design tension in accreditation framework development.
- **Section 5.4.5 (Evaluator training):** The proposal for a "digital assessment specialist"evaluator role is sensible but raises practical questions. HEEACT's evaluator pool is composed of volunteer academics; adding a new specialist requirement may reduce the already-limited pool of willing evaluators. The paper should discuss how this specialist role would be incentivized and sustained.
- **Section 5.5 (International comparison):** The comparison with Singapore is apt, but the paper overstates Singapore's relevance as a comparator. Singapore's higher education system has fewer than 10 autonomous universities, compared to Taiwan's 150+. Scale differences fundamentally affect implementation feasibility. The comparison with TEQSA (Australia, ~170 providers) is more relevant for scale considerations.

Questions for Authors

1. What is your assessment of the realistic timeline for fourth-cycle design decisions? When does the window for influencing the framework close?
2. How do you propose addressing the resource paradox —that institutions most in need of improved assessment are least capable of implementing AI systems?
3. Have you consulted with HEEACT staff or MOE officials about the feasibility of the proposed Core Indicator revisions? If not, how do you plan to validate the policy recommendations?
4. How do you address the language barrier —the fact that most agentic AI systems are trained predominantly on English-language data and may perform poorly in Chinese-language educational assessment?

5. What role, if any, do you envision for students and student unions in the governance of AI-augmented assessment? The paper's ethics section discusses student rights but not student participation in governance.

Recommendation: Minor Revision

The paper demonstrates genuine domain expertise and offers specific, actionable policy recommendations. The weaknesses identified are significant but addressable without restructuring the paper's core argument. I recommend minor revision with particular attention to: (a) the fourth-cycle timeline specificity, (b) the political economy of HEEACT-MOE governance, (c) the resource paradox of the demographic crisis, and (d) the language/cultural dimension of AI deployment in a Chinese-language educational context.

Reviewer 3 (R3): Cross-Disciplinary Expert

Overall Assessment

This paper makes a welcome contribution to the emerging literature on AI and higher education quality assurance by focusing specifically on agentic AI —a technological category that is qualitatively distinct from the generative AI tools that dominate current discourse. The distinction between passive, generative, and agentic AI systems is important and undertheorized in the educational literature; this paper advances that conversation meaningfully. The four-level taxonomy of AI agency in education (Section 2), while not without problems, provides a useful vocabulary for differentiating the capabilities and risks associated with different classes of AI systems. The ethical analysis in Section 6 is the paper's strongest section —rigorous, nuanced, and genuinely attentive to the risks of techno-solutionism.

However, the paper has notable weaknesses in its treatment of the AI ethics literature, its engagement with the international scholarship on AI in education, and its tendency —despite self-conscious hedging —to present a vision of agentic AI assessment that is more technologically optimistic than the current evidence warrants. The taxonomy, while useful, lacks clear operationalization criteria. And the paper's engagement with algorithmic bias, while substantive, does not go deep enough into the specific mechanisms through which bias operates in educational AI systems.

Strengths

1. **The agentic AI taxonomy (Section 2) is a genuine contribution.** The four-level taxonomy —reactive tool-based AI, autonomous single-agent AI, adaptive multi-agent AI, and AI assessment ecosystem — provides conceptual clarity that is missing from much of the current literature. The distinction between generative AI (which produces outputs on demand) and agentic AI (which plans, adapts, and acts with delegated authority) is crucial for policy analysis, and this paper makes it clearly.
2. **Section 6.3 (Risks unique to agentic AI) is outstanding.** The identification of five agentic-specific risks —autonomy amplification, iterative reasoning unpredictability, persistent memory bias compounding, multi-agent coordination failures, and goal misalignment —goes beyond the generic AI ethics discourse to identify risks that are specific to the technological architecture being proposed. This is exactly the kind of differentiated risk analysis that the field needs. The "persistent memory bias compounding"risk, in particular, is an original and important contribution —the idea that early assessment errors can create self-reinforcing bias trajectories through longitudinal AI profiles.

3. **The paper avoids crude techno-solutionism.** Section 4.5 (The Role of Human Judgment) and Section 7.2 (What This Paper Does NOT Claim) explicitly reject technology determinism and insist on the irreducibility of human judgment. The "centaur model"metaphor (Section 4.5), drawn from Kasparov' s chess analysis, is apt and effectively communicated.

4. **The three-tier governance framework (Section 6.4) is well-structured.** The progression from national governance (MOE/HEEACT standards), through institutional governance (AI Assessment Ethics Committees, Student Data Bill of Rights), to technical governance (algorithmic transparency, bias testing, data minimization, interoperability) is logical and comprehensive. The Student Data Bill of Rights with its five core rights (consent, access, correction, deletion, portability) is a practical and important proposal.

5. **The paper correctly identifies the tension between formative and summative assessment functions (Section 4.3.6).** This is a genuinely difficult design problem: if the same AI system supports learning and generates evidence for accreditation, conflicts of interest are inevitable. The paper names this problem clearly, even if it does not fully resolve it.

Weaknesses

1. **The agentic AI taxonomy lacks operational criteria for level differentiation.** The four levels are described in terms of capabilities (planning, adaptation, tool use, etc.), but the boundaries between levels are not operationally defined. What observable features distinguish a Level 2 (autonomous single-agent) system from a Level 3 (adaptive multi-agent) system? Without clear boundary conditions, the taxonomy risks being applied inconsistently. The paper should either provide operational definitions or acknowledge that the levels represent a continuum rather than discrete categories.

2. **The paper's engagement with the algorithmic bias literature is broad but shallow.** Baker and Hawn (2022) and Gandara et al. (2024) are cited as evidence of bias in educational AI, and the paper correctly notes the relevance of these findings to Taiwan's marginalized populations. However, the paper does not engage with the specific mechanisms of bias —how training data composition, feature engineering choices, proxy variables, and outcome variable definition each contribute to differential performance. For a paper proposing to embed AI in consequential assessment decisions, a deeper engagement with the technical mechanisms of bias is necessary. The paper should also address whether the bias mitigation strategies it proposes (disaggregated analysis, diverse training data, equity impact assessments) are actually effective —the evidence on bias mitigation in educational AI is more mixed than the paper implies.

3. **The paper does not adequately address the problem of construct validity for AI-generated competency assessments.** The paper argues that agentic AI can assess complex competencies like critical thinking, creative problem-solving, and ethical reasoning (Section 3.3, Section 4.3.5). But establishing that an AI system is actually measuring these constructs —rather than proxies for them —is a fundamental validity challenge that the paper largely sidesteps. Mislevy et al. (2003) and Evidence-Centered Design are cited, but the paper does not engage with the substantial psychometric literature on the difficulty of validating assessments of complex competencies even with human raters. The validity problem is arguably harder for AI systems, not easier.

4. **The international literature review is selective.** The paper engages well with some strands of the AI-in-education literature (Bearman & Ajjawi, 2023; Swiecki et al., 2022; Banihashem et al., 2025) but omits significant work. Holmes et al. (2022, "Ethics of AI in Education") and Williamson (2017, "Big Data in Education") are conspicuous absences. The UNESCO (2023) guidance is cited but the

more critical perspectives within the UNESCO framework —including concerns about AI exacerbating global educational inequalities —are not fully engaged. The Selwyn (2019) reference in Section 5.4.5 is tokenistic; Selwyn's broader critique of educational technology hype deserves more substantive engagement.

5. **The paper oscillates between careful hedging and strongly normative claims.** Section 7.2 explicitly disclaims empirical claims, but the conclusion (Section 8) shifts to language like "the paradigm shift is coming whether Taiwan prepares for it or not"and "do not wait for perfect evidence."This tension between epistemic caution and advocacy creates a tonal inconsistency that may undermine the paper's credibility with skeptical readers. A theoretical paper should maintain a consistent epistemic register.

6. **The treatment of student agency in the agentic paradigm is underdeveloped.** The paper envisions students as "co-creators"of assessment evidence (Section 4.3.3) and describes Wei-Lin annotating disagreements with AI assessments (Section 4.4). But the paper does not adequately address the power asymmetry inherent in asking students to "co-create"evidence with a system that has persistent memory, institutional backing, and algorithmic authority. When a student disagrees with an AI assessment, whose judgment prevails? Under what conditions? Through what institutional mechanism? The paper' s vision of student co-creation risks being naive about the power dynamics involved.

Specific Comments

- **Section 2 (AI taxonomy):** The reference to "Inside Higher Ed (2026, January)"reporting on an AI agent "Einstein"passing university courses autonomously is striking but unverified. If this is a real news report, the citation should be complete; if it is speculative, it should be clearly flagged as such. Using unverified future-dated sources in a theoretical paper undermines credibility.
- **Section 4.3.5 (Evidence type):** The concept of "stealth assessment"(Shute & Ventura, 2013) is introduced approvingly but deserves more critical engagement. Stealth assessment —extracting evidence from natural interactions without student awareness —raises significant consent issues that are in tension with the paper's own autonomy principles in Section 6.1.1. The paper should address this tension directly.
- **Section 6.1.2 (Beneficence):** The acknowledgment that the evidence base is "still emerging and largely inconclusive"(UNESCO, 2023) is important but should be placed more prominently. Currently it appears in the ethics section rather than in the paper's core argument about agentic AI capabilities.
- **Section 6.4 (Governance framework):** The proposal for AI Assessment Ethics Committees modeled on IRBs is sensible but the paper should acknowledge the well-documented challenges with IRB effectiveness —bureaucratic delay, inconsistent standards, limited expertise —that might afflict analogous committees in the AI context.

Questions for Authors

1. How do you establish construct validity for AI-generated assessments of complex competencies like critical thinking and ethical reasoning? What psychometric framework would you propose?
2. The paper advocates "stealth assessment"(Section 4.3.5) but also insists on informed consent (Section 6.1.1). How do you reconcile these positions?
3. What evidence supports the effectiveness of the bias mitigation strategies you propose (disaggregated analysis, diverse training data, equity impact assessments)? Are these strategies sufficient to address

the structural nature of algorithmic bias in education?

4. When a student disagrees with an AI assessment (as in the Wei-Lin scenario), whose judgment prevails? Under what institutional and procedural conditions?
5. How do you respond to the critique (Selwyn, 2019; Williamson, 2017) that educational technology scholarship systematically overestimates the transformative potential of new technologies while underestimating the persistence of institutional structures?

Recommendation: Major Revision

The paper makes a genuine contribution through its agentic AI taxonomy and its differentiated ethical analysis, but requires significant strengthening in three areas: (a) construct validity for AI-generated competency assessments, (b) deeper engagement with the mechanisms of algorithmic bias and the evidence on bias mitigation, and (c) more consistent epistemic register throughout the paper. The treatment of student agency and the tension between stealth assessment and informed consent also require resolution.

Reviewer 4: Devil's Advocate

Fundamental Challenges

1. **The paper's central analogy —Kuhnian paradigm shift —is fundamentally flawed.** Kuhn's theory describes how scientific communities abandon one explanatory framework for another when anomalies accumulate. Quality assurance is not a scientific paradigm; it is an administrative and regulatory regime. Administrative regimes change through political negotiation, institutional inertia, and incremental reform —not through revolutionary gestalt shifts. By clothing a policy reform proposal in Kuhnian language, the paper imports a false sense of inevitability ("the paradigm shift is coming whether Taiwan prepares for it or not,"Section 8) that is not warranted by the actual dynamics of institutional change in higher education. Quality assurance frameworks in every major national system have evolved incrementally over decades; none has undergone anything resembling a Kuhnian revolution. The paper's own recommendation —Scenario B, incremental framework evolution —contradicts the revolutionary framing.
2. **The ADAPT framework is not a framework; it is an acronym imposed on the paper's table of contents.** Strip away the acronym and what remains? A paper that: (a) defines agentic AI, (b) diagnoses current QA limitations, (c) proposes new assessment concepts, (d) evaluates policy options, and (e) discusses ethics. This is perfectly standard paper structure. Calling it the "ADAPT framework" does not make it a theoretical contribution. A genuine framework would have internal logic —the components would be related to each other in ways that generate predictions, explain relationships, or constrain possibilities. ADAPT does none of these things. It is a mnemonic device masquerading as a conceptual contribution.
3. **The paper has no empirical content whatsoever.** It acknowledges this (Section 7.2, 7.3), but the acknowledgment does not mitigate the problem. The paper proposes a radical transformation of Taiwan' s assessment infrastructure based on: (a) theoretical reasoning about what agentic AI could do, (b) a single cited empirical study (Kestin et al., 2025) from a completely different context, and (c) optimistic projections from industry reports (Gartner, 2025) and preprints. A Q1 journal in higher education — Studies in Higher Education, Higher Education, or Assessment & Evaluation in Higher Education —

expects papers to be grounded in evidence, not in speculation about technology that does not yet exist at the scale the paper envisions.

4. **The paper systematically understates the gap between current AI capabilities and the vision it describes.** The Wei-Lin scenario (Section 4.4) describes an AI system that: continuously monitors learning behaviors across all courses, detects subtle competency plateaus, generates diagnostic analyses of probable causes, recommends specific interventions, coordinates multi-agent assessment across multiple dimensions during a capstone project, and produces "evidence narratives"that synthesize four years of longitudinal data. No such system exists. The closest approximations —learning analytics dashboards, adaptive testing systems, AI tutors —operate at orders of magnitude less sophistication. The paper occasionally acknowledges this gap ("The gap between current capability and this integrated vision remains significant,"Section 4.4) but then proceeds as if the gap were merely an engineering detail rather than a fundamental question about feasibility.

5. **The paper ignores the most likely failure mode: institutional performativity.** If HEEACT incorporates AI assessment metrics into accreditation standards, the most probable institutional response is not genuine transformation but performative compliance —purchasing an AI system, generating the required reports, and continuing business as usual. This is exactly what happened with outcomes-based education: institutions adopted the language of OBE, created the required documentation, and largely continued teaching as before (Lin et al., 2020, as cited by the paper itself). The paper does not address why AI adoption would avoid this pattern of surface compliance.

Weakest Arguments

1. **The claim that agentic AI addresses the "agency limitation"(Section 4.3.3) is circular.** The paper argues that the current paradigm suffers from information asymmetry because institutions curate their own evidence. The proposed solution is AI agents that "independently access learning management systems, assignment repositories, and institutional databases."But who deploys these AI agents? The institutions. Who configures them? The institutions. Who determines what data they access and how they report it? The institutions. The agency limitation is not solved by replacing human-authored reports with institutionally-deployed AI-authored reports; it is merely relocated. True independence would require HEEACT to deploy its own AI agents within institutions —a scenario the paper does not propose, presumably because it would be politically untenable.

2. **The "seven dimensions of paradigm shift"conflate what is technologically possible with what is educationally desirable.** The paper assumes that continuous monitoring is better than periodic assessment, that individual tracking is better than aggregate measurement, and that real-time feedback is better than retrospective reporting. None of these assumptions is self-evident. Periodic assessment with human reflection may produce deeper institutional learning than continuous automated monitoring. Aggregate measures may protect individual privacy in ways that individualized tracking cannot. Retrospective analysis may enable the kind of contextual, meaning-making interpretation that real-time processing forecloses. The paper treats each dimension as a unidirectional improvement rather than a trade-off.

3. **The ethical governance proposals are aspirational but untested.** The paper proposes AI Assessment Ethics Committees, a Student Data Bill of Rights, algorithmic transparency requirements, and bias testing protocols. Each of these is reasonable in principle, but none has been tested in a higher education assessment context. The paper presents them as solutions without evidence that they would actually work. IRBs —the model for the proposed Ethics Committees —are widely criticized for bureaucratic inefficiency and inconsistent standards. Algorithmic transparency is technically challenging

for complex AI systems. Bias testing protocols in other domains have produced mixed results. The governance framework is a wish list, not a demonstrated solution.

4. **The paper's treatment of the "centaur model"(Section 4.5) is idealized.** The chess analogy is elegant but misleading. In chess, the rules are fixed, the objective is clear, and performance is unambiguously measurable. Educational assessment has none of these properties. The values are contested, the objectives are multiple and often conflicting, and performance is inherently ambiguous. The conditions under which centaur teams outperform humans or AI alone in chess do not obviously transfer to educational assessment, where the "game"is not well-defined.

5. **The conclusion's call to action ("do not wait for perfect evidence") contradicts the paper's own epistemic commitments.** Section 7.2 explicitly states that the paper "does not provide empirical evidence that agentic AI improves learning outcomes."Section 7.3 acknowledges that the evidence base is "still emerging and largely inconclusive."Yet Section 8 urges Taiwan to "act within the window, with the evidence at hand."This is a policy advocacy move, not an academic conclusion. A theoretical paper should conclude with a research agenda, not a call to action based on evidence it acknowledges does not exist.

Missing Counter-Arguments

1. **Goodhart's Law applied to learning trajectories.** The paper identifies goal misalignment (Section 6.3) but does not apply it to its own core proposal. If "learning trajectories"replace graduation rates as the unit of assessment, institutions and students will optimize for trajectory metrics —producing impressive-looking trajectories that may not correspond to genuine learning. The very continuity and granularity that the paper celebrates make the system more vulnerable to gaming, not less.

2. **The totalizing surveillance implications of continuous monitoring.** The paper mentions surveillance risk but does not engage with the deeper argument that continuous AI monitoring of student learning behaviors is fundamentally incompatible with the intellectual freedom that universities exist to protect. A university is not a factory where every worker's productivity should be monitored; it is a space where students should be free to explore, fail, waste time, pursue tangents, and develop in ways that are not captured by competency trajectories.

3. **The political economy of AI vendors.** The paper does not discuss who would build and maintain the agentic AI systems it envisions. The most likely answer is commercial technology companies with profit motives that may not align with educational values. Vendor lock-in, data extraction, and the commercialization of learning evidence are significant risks that the paper does not address.

4. **The possibility that the "paradigm shift"is just hype.** The paper takes at face value the claims of AI industry reports (Gartner, 2025), preprints (Tao et al., 2026), and a single RCT (Kestin et al., 2025). It does not engage with the extensive literature on technology hype cycles in education —the repeated pattern of revolutionary promises followed by disappointing implementations. Interactive whiteboards, MOOCs, learning management systems, and virtual reality were all going to "transform" education. None did. What makes agentic AI different?

Strongest Counter-Evidence

The paper's thesis would be most seriously challenged by evidence demonstrating that: (a) existing learning analytics and AI tutoring systems have not produced sustained improvements in learning outcomes when deployed at scale in diverse institutional contexts; (b) continuous monitoring of student behavior leads to performative compliance rather than genuine learning improvement; (c) the technical challenges of validating

AI-generated competency assessments (construct validity, reliability, fairness) are intractable rather than merely unsolved; or (d) faculty and student resistance to AI-mediated assessment is structural rather than transitional —rooted in legitimate values about education's purposes that technological capability cannot override.

Evidence along lines (a) and (d) already exists in the literature but is not adequately engaged by the paper. The learning analytics field has produced disappointing results at scale (Ferguson & Clow, 2017; Viberg et al., 2018), and faculty resistance to educational technology is well-documented as values-based rather than merely attitudinal (Selwyn, 2019).

Recommendation: Major Revision

The paper is intellectually ambitious and well-written, but its core argument rests on assumptions that are insufficiently examined and evidence that is insufficiently robust. The Kuhnian framing needs to be either substantially justified or replaced. The ADAPT framework needs to demonstrate genuine analytical utility beyond acronymic organization. The gap between current AI capabilities and the paper's vision needs to be addressed more honestly. And the paper needs to engage seriously with the counter-arguments identified above —not merely acknowledge them in a limitations section but integrate them into the core argument. If the authors can accomplish this, the paper could make a significant contribution. In its current form, it does not meet the standard of a Q1 journal.

Reviewer 5 (EIC): Editor-in-Chief

Editorial Summary

This paper tackles a topic of genuine importance and timeliness: how should higher education quality assurance systems respond to the emergence of agentic AI? The focus on Taiwan's HEEACT system provides welcome specificity in a literature that too often deals in generalities. The paper is ambitious in scope — spanning conceptual taxonomy, paradigm shift theory, policy analysis, and ethical governance —and is generally well-written, with a confident and engaging prose style that makes complex arguments accessible. All four reviewers acknowledged the paper's intellectual merit and the originality of several contributions, particularly the seven dimensions of paradigm shift (Section 4.3), the agentic-specific risk analysis (Section 6.3), and the concrete fourth-cycle accreditation recommendations (Section 5.4).

However, the reviews converge on several significant concerns. R1 and R4 both challenge the appropriateness of the Kuhnian framework for analyzing what is essentially a policy reform proposal. R1 and R4 both question whether the ADAPT framework constitutes a genuine theoretical contribution or merely organizes the paper's structure. R1 and R3 identify the thin evidence base for agentic AI capabilities as a foundational weakness. R3 and R4 both note the paper's oscillation between epistemic caution and advocacy. R2 raises important domain-specific issues —the political economy of HEEACT-MOE governance, the resource paradox of the demographic crisis, and the language dimension of AI deployment —that the paper must address to be credible with its target audience. R4 raises fundamental challenges about institutional performativity, the idealization of the centaur model, and the omission of technology vendor dynamics.

Key Strengths (Consensus)

1. **The seven dimensions of paradigm shift (Section 4.3)** are recognized by all reviewers as the paper's strongest analytical contribution —clearly defined, well-structured, and genuinely useful for policy analysis.

2. **The agentic-specific risk analysis (Section 6.3)** is identified by R1, R3, and R4 as an original and important contribution that advances the ethical discourse beyond generic AI ethics.
3. **The fourth-cycle accreditation recommendations (Section 5.4)** are praised by R2 as specific, actionable, and politically astute —the kind of concrete proposals that could actually influence policy.
4. **The paper's intellectual honesty (Sections 7.2 and 7.3)** —its explicit delineation of what it does not claim and its transparent acknowledgment of limitations —is noted by R1 as exemplary.
5. **The writing quality** is consistently high throughout, making a complex, multi-framework argument accessible and engaging. The prose avoids jargon while maintaining precision.

Key Concerns (Consensus)

1. **The Kuhnian framework application is insufficiently justified** (R1, R4). The extension of Kuhn's theory from natural science to administrative/regulatory contexts requires explicit argument that the paper does not provide. The paper's own recommendation (Scenario B, incremental evolution) is in tension with the revolutionary framing.
2. **The ADAPT framework's status as a theoretical contribution is questionable** (R1, R4). It may be a re-labeling of the paper's research questions (R1) or an acronym imposed on standard paper structure (R4). The derivation logic and internal relational structure need to be made explicit.
3. **The evidence base for agentic AI capabilities is thin** (R1, R3, R4). The paper relies heavily on preprints, industry reports, and a single empirical study from an unrepresentative context. For a paper making claims about paradigm-level transformation, the evidentiary foundation must be stronger.
4. **The gap between current AI capabilities and the paper's vision is understated** (R3, R4). The Wei-Lin scenario describes a system far beyond current technological capability, and the paper does not adequately address the feasibility question.
5. **The paper oscillates between epistemic caution and advocacy** (R3, R4). The careful hedging in Sections 7.2-7.3 is undermined by the call-to-action rhetoric in Section 8. The paper needs a consistent epistemic register.
6. **Domain-specific gaps weaken the policy analysis** (R2). The political economy of HEEACT-MOE governance, the resource paradox, the language dimension, and institutional performativity dynamics are insufficiently addressed.

Editorial Decision: Major Revision

Rationale

This paper addresses a topic of substantial importance and demonstrates intellectual ambition, wide reading, and genuine analytical capability. Several of its contributions —the seven dimensions of paradigm shift, the agentic-specific risk analysis, and the fourth-cycle accreditation recommendations —are valuable and would enrich the literature. The paper is well-written and engages its target audience effectively.

However, the paper's two primary theoretical claims —the Kuhnian paradigm shift framing and the ADAPT framework —both require significant strengthening. The Kuhnian framework is applied without adequate justification of its applicability to administrative/regulatory contexts, and the ADAPT framework has not demonstrated the internal logic and derivation rigor needed to qualify as a genuine theoretical contribution. These are not superficial problems; they affect the paper's central intellectual architecture.

Additionally, the evidence base is thin for a paper making claims about paradigm-level transformation. The gap between current AI capabilities and the paper's vision needs to be addressed more honestly and systematically. The paper's tonal oscillation between careful hedging and policy advocacy creates a credibility problem that must be resolved.

I am recommending Major Revision rather than Reject because the paper's strengths are genuine, its topic is important, and the identified weaknesses are addressable. However, the revision must be substantive, not cosmetic. Specifically, the authors must: (1) either provide rigorous justification for the Kuhnian framing or adopt a less freighted theoretical lens; (2) demonstrate that the ADAPT framework has internal relational logic beyond acronymic organization; (3) strengthen the evidence base or calibrate claims to the available evidence; (4) address the domain-specific gaps identified by R2; (5) engage seriously with the counter-arguments raised by R4; and (6) adopt a consistent epistemic register throughout. I look forward to reviewing a revised manuscript that realizes the genuine potential of this work.

Revision Roadmap

Required Revisions (Must Address)

#	Issue	Source	Section(s)	Priority
1	Justify applicability of Kuhnian framework to administrative/regulatory QA contexts, or adopt alternative theoretical framing. Engage with critiques of extending Kuhn beyond natural sciences. Address the tension between "paradigm revolution" framing and Scenario B (incremental evolution) recommendation.	R1, R4	4.1, 8	Critical

#	Issue	Source	Section(s)	Priority
2	Demonstrate ADAPT framework's derivation logic and internal relational structure. Show why these five components (and not others) are neces-sary/sufficient. Distinguish ADAPT from a re-labeling of research questions or paper structure.	R1, R4	4.2	Critical
3	Strengthen evidentiary base for agentic AI capabilities. Add more robust empirical evidence beyond preprints, industry reports, and a single RCT. If evidence is genuinely thin, recalibrate claims accordingly with explicit hedging.	R1, R3, R4	2, 4.3, 4.4	Critical

#	Issue	Source	Section(s)	Priority
4	Address the gap between current AI capabilities and the Wei-Lin scenario more honestly. Specify which elements are near-term feasible, which are medium-term plausible, and which are long-term speculative.	R3, R4	4.4	High
5	Resolve tonal inconsistency between epistemic caution (Sections 7.2-7.3) and advocacy rhetoric (Section 8). Adopt consistent epistemic register. The conclusion should not claim urgency based on evidence the paper acknowledges does not exist.	R3, R4	7.2, 7.3, 8	High

#	Issue	Source	Section(s)	Priority
6	Address political economy of HEEACT-MOE governance in policy recom-mendations. Discuss who has decision-making authority, budget allocation processes, and coordination challenges.	R2	5.3, 5.4	High
7	Address resource paradox: institutions most affected by demographic crisis are least capable of AI adoption. Propose specific equity mechanisms (e.g., shared infrastructure, consortium models, MOE funding).	R2, R4	3.4, 5.3, 5.4	High
8	Address institutional performativity: why would AI adoption avoid the pattern of surface compliance documented for OBE? What design features prevent performative rather than genuine adoption?	R4	4.3, 5.1	High

#	Issue	Source	Section(s)	Priority
9	Engage with construct validity challenge for AI-generated competency assessments. What psychometric framework establishes that AI is measuring critical thinking, ethical reasoning, etc., rather than proxies?	R3	4.3.5, 4.3.2	High
10	Fix duplicate Table 1 numbering (Section 3.2 and Section 4.3 both have “Table 1”). Renumber all tables sequentially.	R1	3.2, 4.3	High

Recommended Revisions (Should Consider)

#	Issue	Source	Section(s)	Priority
11	Address language/cultural dimension of AI deployment in Chinese-language educational contexts. Acknowledge that most agentic AI systems are trained on English-language data and discuss implications.	R2	2, 4.3	Medium
12	Make Bardach's eightfold path visible in the policy analysis. Either systematically apply the eight steps or downgrade Bardach from "core methodological framework"to "analytical inspiration."	R1	5	Medium

#	Issue	Source	Section(s)	Priority
13	Deepen engagement with algorithmic bias mechanisms — not just that bias exists, but how it operates technically (training data, feature engineering, proxy variables, outcome definition). Assess effectiveness of proposed mitigation strategies.	R3	6.1.3, 6.1.4	Medium
14	Resolve tension between “stealth assessment” advocacy (Section 4.3.5) and informed consent requirements (Section 6.1.1).	R3	4.3.5, 6.1.1	Medium
15	Differentiate faculty resistance by institutional type (research vs. teaching-focused) and by resistance type (philosophical vs. practical). Propose differentiated strategies.	R2	3.4, 5.4.5	Medium

#	Issue	Source	Section(s)	Priority
16	Discuss AI vendor dynamics—who builds these systems, commercial incentives, vendor lock-in risks, data extraction concerns.	R4	5, 6	Medium
17	Add operational criteria for level differentiation in the agentic AI taxonomy. Define boundary conditions between levels or acknowledge the continuum nature.	R1, R3	2	Medium
18	Operationalize transition criteria between the three implementation phases. What specific evaluation outcomes trigger progression from Phase 1 to Phase 2?	R1	5.3	Medium
19	Address student agency and power asymmetry in AI co-creation scenarios. When student and AI disagree, whose judgment prevails and through what mechanism?	R3	4.3.3, 4.4	Medium

#	Issue	Source	Section(s)	Priority
20	Engage more substantively with Selwyn (2019) and the broader critique of educational technology hype cycles. Address: what makes agentic AI different from previous "transformative" technologies?	R3, R4	1, 7	Medium
21	Tighten the abstract to 250-300 words (currently approximately 450 words).	R1	Abstract	Low
22	Provide more nuanced characterization of UCAN's evolution and current capabilities.	R2	3.2	Low
23	Verify the "Inside Higher Ed (2026, January)"citation about AI agent "Einstein."If speculative, flag clearly.	R3	2	Low
24	Clarify fourth-cycle accreditation timeline specifics —when does the design window open/close?	R2	5.4	Low

#	Issue	Source	Section(s)	Priority
25	Discuss challenges with IRB effectiveness as a caution for the proposed AI Assessment Ethics Committees.	R3, R4	6.4	Low

Editorial Guidance

The revision should proceed in three stages. First, address the two critical issues: the Kuhnian framing and the ADAPT framework. For the Kuhnian framing, the authors have two viable options: (a) provide a rigorous meta-theoretical justification for applying Kuhn to regulatory contexts, engaging with critics and specifying what counts as an "anomaly"and "crisis"in an administrative regime; or (b) replace the Kuhnian framing with a less contentious theoretical lens —institutional theory (DiMaggio & Powell, 1983), policy learning frameworks (Sabatier & Jenkins-Smith, 1993), or critical realism —that can support the same analytical work without the metaphysical baggage of incommensurability and revolution. For the ADAPT framework, the authors should articulate the generative logic: how do the five components relate to each other? What does the framework predict or explain that a simple list of topics would not? If the framework's value lies in its integrative function, that function should be demonstrated through application, not merely asserted.

Second, address the evidentiary gap. The authors should either: (a) conduct a more systematic review of the empirical evidence on AI in educational assessment, including negative and null results; or (b) explicitly reframe the paper as a speculative/futures-oriented analysis with appropriate epistemic markers throughout. The current hybrid —claiming theoretical rigor while relying on speculative evidence —satisfies neither standard.

Third, integrate the domain-specific feedback from R2 and the counter-arguments from R4 into the body of the paper, not merely into a limitations section. The strongest version of this paper would anticipate and engage with its own best critiques rather than deferring them to a separate section. The institutional performativity argument, the resource paradox, the vendor dynamics, and the technology hype cycle critique should each be addressed where they are most analytically relevant —in Sections 4, 5, and 6 respectively.

The revised manuscript should aim for a consistent epistemic register: ambitious in its theoretical vision, honest about the current evidentiary limitations, and precise in distinguishing between what is demonstrated, what is plausible, and what is speculative. This paper has the potential to make a significant contribution to the literature on AI and higher education quality assurance. Realizing that potential requires the authors to hold their own work to the same rigorous standard they apply to the assessment paradigm they seek to transform.