Artificial Intelligence Education Studies
Volume 1, Issue 3, 2025, Page 43–55
https://doi.org/10.6914/aiese.010304

Artificial Intelligence Education Studies

From Traces to Teaching: A Socio-Technical Framework for Process-Based Assessment in an Age of Distributed Cognition

Alexander LEE

Hong Kong Journal of Social Sciences Press

Abstract

Generative AI severs the link between polished products and genuine learning, exposing the limits of outcome-only assessment. This paper advances a socio-technical framework for AI-enabled process-based assessment (PBA) that reframes evaluation as continuous diagnosis embedded in learning. A five-stage pipeline—task → trace → model → feedback → validation—aligns pedagogical intent with instrumentation of interaction, discourse, and multimodal evidence, and treats the human AI pair as the unit of analysis within a distributed cognition perspective. Methodologically, the framework maps trace types to appropriate model families (e.g., sequential pattern mining, HMMs, NLP) while requiring explainability so insights are actionable. For practice, it specifies teacher AI orchestration roles that preserve human judgment and defines governance protocols for privacy, fairness, transparency, and cultural responsiveness. The result is a principled route to assess complex problem solving with integrity in the age of generative AI.

Keywords Process-Based Assessment; Distributed Cognition; Digital Traces; Explainable AI; Teacher AI Orchestration

1 Introduction: The Assessment Crisis and the Process-Based Imperative

1.1 The Collapse of Outcome-Based Assessment in the Generative AI Era

Complex problem solving (CPS) is widely recognized as a cornerstone of 21st-century education, essential for navigating the ambiguous, dynamic, and interconnected challenges of modern life. Its significance is formally acknowledged through its inclusion in large-scale international evaluations, such as the OECD's Programme for International Student Assessment (PISA), which defines it as a critical competency for future readiness. Unlike well-structured problems with clear solution paths, CPS tasks are often ill-defined, demanding iterative strategies, metacognitive regulation, and effective collaboration to achieve resolution.

Despite this consensus, the predominant assessment paradigm, Outcome-Based Assessment (OBA), remains fundamentally misaligned with the nature of CPS. OBA focuses on the final product—a test score, a project deliverable, or the correctness of a final submission—while treating the rich, dynamic process of exploration, strategy formulation, and adaptation as an invisible "black box". This methodological gap has always limited educators' ability to provide the nuanced, formative feedback required to develop these sophisticated skills. However, the recent and rapid proliferation of generative AI has transformed this long-standing limitation into an existential crisis for assessment integrity.

Generative AI tools can produce fluent, sophisticated, and often correct final products with minimal authentic student effort, fundamentally severing the link between the observable outcome and the student's underlying knowledge or skill. When a final essay, a piece of code, or a project report can no longer be trusted as a reliable signal of learning, the validity of OBA collapses. This technological disruption has created what many have termed an "assessment crisis," forcing a re-evaluation of the very foundations of educational measurement. More profoundly,

© The Authors 2025. Article History Received: June 9, 2025; Accpted: July 5, 2025; Published: September 30, 2025. ***Artificial Intelligence Education Studies***, ISSN 3079-8086 (print), ISSN 3079-8094 (online), a Quarterly, founded on 2025, published by Creative Publishing Co., Limited. Email: wtocom@gmail.com, https://ai-es.org, https://cpcl.hk. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits non-commercial use, sharing, adaptation, distribution, and reproduction in any medium or format, provided the original author(s) and source are credited, and derivative works are licensed under the same terms.

it has inverted the value proposition of assessment. Previously, the product was the trusted signal and the process was the noisy, hard-to-measure context. In the current landscape, the product is potentially untrustworthy, making the process the only remaining reliable signal of authentic student engagement, critical thinking, and learning. This shift makes a transition to process-focused evaluation not merely a pedagogical improvement but a methodological necessity for restoring validity and meaning to assessment. The initial reaction to this crisis, focused on detecting AI-generated text and preserving academic integrity, addressed only a surface-level symptom. The deeper challenge is not simply to prevent cheating, but to fundamentally rethink what constitutes valid evidence of learning when the final artifact is no longer sufficient.

1.2 The Methodological Necessity of Process-Based Assessment (PBA)

Process-Based Assessment (PBA) emerges as the necessary paradigm to address this crisis. Defined as an approach that analyzes the how and why of student performance by examining the sequence of actions, decisions, and strategies employed during a task, PBA makes the thinking process visible. This focus is deeply rooted in established learning theories, from constructivism, which values the process as much as the product, to models of self-regulated learning (SRL), which conceptualize learning as a cyclical process of goal-setting, strategy enactment, and adaptation, as articulated in seminal frameworks like that of Winne and Hadwin.

For decades, the widespread adoption of PBA was hindered by practical constraints; manual analysis of process data was too resource-intensive to be feasible at scale. Today, a powerful confluence of three factors has made scalable PBA a reality. First is the urgent pedagogical need to assess the complex skills demanded by the 21st century, a need now amplified by the validity challenges posed by generative AI. Second is the ubiquitous availability of rich data, as digital learning environments from virtual labs to collaborative platforms generate massive streams of "digital traces" —fine-grained logs of student interactions. Third is the analytical capability of Artificial Intelligence, which provides the computational power to analyze these vast and complex datasets, identify latent patterns, and infer cognitive and affective states from behavior. This convergence creates an unprecedented opportunity to elevate PBA from a niche research method to a sustainable and scalable educational practice, offering a robust response to the crisis catalyzed by AI itself.

1.3 Argument and Contributions: A Principled Framework for Assessment Reimagined

This paper argues that AI-enabled PBA offers a transformative approach to assessing CPS, reframing assessment as a continuous, diagnostic process embedded within learning activities themselves. This synthesis resolves the long-standing tension between the pedagogical ideal of understanding the learning process and the practical constraints of traditional assessment, providing a robust path forward in an AI-mediated world. To operationalize this vision, this paper makes four key contributions.

First, a Socio-Technical Design Framework: We propose a principled five-stage pipeline (task→trace→model→feedback→validat
for developing trustworthy AI-PBA systems that prioritizes pedagogical goals over purely technical considerations. This framework is explicitly grounded in socio-technical systems theory to ensure a holistic and responsible design process.

Second, a Modernized Evidentiary Basis: We provide methodological guidance for capturing and abstracting not only traditional digital traces (interaction, discourse, multimodal) but also the new, complex traces of human-AI interaction. We argue for a theoretical shift in the unit of analysis from the individual student to the human-AI distributed cognitive system.

Third, a Protocol for Trustworthy Implementation: We articulate an integrated approach to validity, robustness, and ethical governance, featuring an operational compliance checklist and a novel focus on ensuring cultural responsiveness in AI-based assessment, moving beyond narrow statistical definitions of fairness.

Fourth, a Vision for Human-AI Orchestration: We present a practical model for the collaborative roles of AI and human educators, positioning the teacher as a "classroom orchestrator" who leverages AI-driven insights to make informed pedagogical decisions, thereby augmenting rather than replacing human expertise.

2 A Socio-Technical Framework for Principled AI-PBA Design

To move AI-enabled PBA from a collection of disparate methods to a principled and responsible practice, a systematic design framework is required. The five-stage pipeline proposed here is not merely a technical procedure but a

socio-technical governance model. It is grounded in Socio-Technical Systems (STS) theory, which conceptualizes any work system as comprising interdependent social and technical subsystems, arguing that system success can only be achieved through their "joint optimization". In the context of AI-PBA, the "social" subsystem encompasses pedagogical goals, classroom culture, teacher and student roles, and learning objectives, while the "technical" subsystem includes the AI models, data infrastructure, and digital learning environment. The framework is explicitly designed as a bulwark against "technological solutionism"—the tendency to develop an AI tool and then search for an educational problem to solve. By structuring the process with educational goals as the primary driver, it ensures that AI serves pedagogy, not the other way around. Each stage represents a critical decision point where pedagogical values, human factors, and ethical considerations are embedded, reframing the development of AI-PBA as an act of responsible design from the outset.

2.1 The Five-Stage Design Pipeline: From Pedagogical Primacy to Integrated Validation

The proposed framework is a cyclical governance structure where each stage represents a point of deliberate interweaving between the social and technical subsystems, operationalizing the principle of joint optimization.

2.1.1 Stage 1: Pedagogical Primacy in Task Design

The process begins not with data or algorithms, but with pedagogy. The design of the learning task is paramount, as it must be sufficiently complex and open-ended to elicit the target CPS processes that are to be assessed. This stage forces the social subsystem (pedagogy) to define the requirements for the technical system. Tasks should not be arbitrary but grounded in established theoretical frameworks, such as the PISA 2015 model, which delineates problem-solving processes like "exploring and understanding" and "planning and executing" alongside collaboration processes like "establishing and maintaining shared understanding". The task environment must be designed to afford students opportunities to plan, experiment, make mistakes, and revise their strategies, as these behaviors constitute the very evidence of learning that the system aims to capture.

2.1.2 Stage 2: Theory-Driven Trace Design and Instrumentation

Once the pedagogical task is defined, the digital environment must be instrumented to capture relevant evidence. This stage involves a deliberate, theory-driven choice of which digital traces to collect. The decision is not to capture everything possible, but to capture what is meaningful. If the goal is to assess self-regulation, traces related to goal-setting tools, help-seeking behavior, and revision history are crucial. If assessing collaboration, discourse data from chat logs is essential. This stage requires a careful mapping between the target psychological constructs (e.g., planning) and their observable digital manifestations (e.g., use of a planning tool, creation of an outline before writing). The choice of what data not to collect is as important as the choice of what to collect, representing a foundational act of privacy governance within the design process.

2.1.3 Stage 3: Principled Selection of AI Model Families

The choice of an AI model is not an arbitrary technical decision but is dictated by the nature of the trace data and the specific assessment question. The model must be fit for the pedagogical purpose. If the objective is to understand the temporal evolution of student strategies, sequential models such as Sequential Pattern Mining (SPM), Hidden Markov Models (HMMs), or Recurrent Neural Networks (RNNs) are appropriate choices. If the goal is to identify distinct, emergent profiles of problem-solvers without prior labels (e.g., "systematic explorers" vs. "rapid guessers"), then unsupervised clustering algorithms are more suitable. This principled selection ensures that the analytical tool aligns with the educational goal, preventing the common pitfall of applying a technologically impressive but pedagogically inappropriate model.

2.1.4 Stage 4: Bridging Analysis to Action via Feedback Channels

Analysis is meaningless without action. This stage focuses on translating the insights derived from the AI models into feedback that can positively impact teaching and learning. This feedback can be delivered through various channels, such as real-time, automated hints for students; diagnostic dashboards for teachers that highlight class-wide patterns or individual struggles; or summative reports for learners and instructors that visualize strategic development over time. This stage also involves defining "actuation rules"—the specific conditions under which feedback is triggered.

For example, an RNN-based risk model might trigger an alert to a teacher only when a student's predicted probability of non-completion exceeds a threshold of 0.7 for two consecutive weeks, ensuring that interventions are timely but not overwhelming.

2.1.5 Stage 5: Integrating Validation Hooks for Methodological Rigor

Validation cannot be an afterthought; it must be integrated throughout the design process. Each stage of the framework should have explicit "validation hooks" to ensure the system's methodological rigor and trustworthiness. For Task Design (Stage 1), the hook is alignment with established theoretical constructs of CPS or SRL. For Trace Design (Stage 2), it is ensuring the captured data are meaningful proxies for the intended cognitive processes, often validated through think-aloud studies. For Model Selection (Stage 3), it involves comparing model outputs against external criteria, such as the ratings of expert human educators. This final stage creates a crucial feedback loop where the performance of the technical system is judged against the goals of the social system, turning the linear pipeline into a continuous cycle of design, research, and refinement.

2.2 Resisting Technological Solutionism through Responsible Design

By embedding pedagogical goals, human factors, and ethical considerations at each decision point, this socio-technical framework serves as a structured defense against technological solutionism. This approach contrasts sharply with technology-first initiatives that often fail because they neglect the complex interdependencies of the educational context. The framework mandates that technology be shaped by learning needs, not the other way around, positioning the development of AI-PBA systems as an act of responsible, human-centered innovation from its inception. This ensures that the resulting systems are not only technically functional but also pedagogically sound, ethically robust, and sustainable within real-world educational ecosystems.

3 The Evolving Evidentiary Basis: From Digital Traces to Distributed Cognition

The raw material for AI-enabled PBA is the digital trace data students leave as they interact with learning environments. A systematic approach to collecting and preparing this data is foundational to the validity of any subsequent analysis. However, the nature of this evidence is evolving rapidly with the integration of generative AI, requiring a theoretical and methodological shift in what constitutes "data" in learning analytics.

3.1 A Modernized Typology of Digital Evidence: Interaction, Discourse, and Multimodal Traces

Digital traces can be categorized into three main types, each offering a different lens on the learning process.

Interaction Traces: These are the most common form of digital trace and include logs of discrete user actions within a digital environment. Examples include clickstreams, navigation paths, tool usage events (e.g., activating a highlighter), code submissions and compilations in a programming environment, and system-generated events like error messages.

Discourse Traces: This category encompasses all forms of textual and verbal data generated by students. It includes chat logs and forum posts in collaborative settings, written explanations of problem-solving strategies, and, in research contexts, transcribed think-aloud protocols where students verbalize their thoughts while performing a task.

Multimodal Traces: Leveraging advances in sensor technology, Multimodal Learning Analytics (MMLA) incorporates data streams beyond keyboard and mouse interactions. This can include visual data from cameras (e.g., facial expressions, posture, eye-tracking), auditory data from microphones (e.g., prosodic features of speech), and physiological data from wearable sensors (e.g., heart rate variability) to infer cognitive load, affect, and engagement.

The selection of which traces to collect involves a strategic compromise. The act of collecting the richest, most detailed data—particularly multimodal data—can be invasive and may alter the process being measured, a phenomenon known as the "observer effect." Furthermore, a practical tension exists between data richness and scalability. Multimodal data offers the most holistic view but is resource-intensive and raises significant privacy concerns, whereas simpler log data is more scalable but may lack crucial contextual or affective information.

3.2 The New Unit of Analysis: The Human-AI Distributed Cognitive System

The increasing use of generative AI tools by students introduces a novel and critical class of trace data that fundamentally alters the object of assessment. The digital record of learning is no longer solely a product of human action but of a human-AI dialogue. This necessitates a theoretical shift in the unit of analysis. We argue that with the integration of generative AI, the student can no longer be modeled as an isolated cognitive actor. Instead, the assessment must focus on the partnership.

To ground this shift, we introduce the theory of Distributed Cognition (DC). Originating in studies of complex, technology-mediated work like aviation and navigation, DC posits that cognition is not an individual phenomenon confined to the mind but is a process distributed across people, artifacts, and the environment. Knowledge and cognitive processes are offloaded onto and shared among both human and non-human components of a system. When a student collaborates with a generative AI, they form such a system. The AI acts as an external memory, a reasoning partner, and a content generator, fundamentally altering the cognitive processes involved in the task.

Therefore, the proper unit of analysis for PBA in the AI era is the human-AI distributed cognitive system. The boundary of the cognitive system being assessed expands to include the AI. This reframes the central assessment question from "What does the student know?" to "How does the student-AI system collaboratively build and validate knowledge?". This theoretical move has profound methodological implications: data collection must capture the entire dialogic loop between human and AI as the fundamental unit of evidence, changing the very ontology of "learning data."

3.3 Methodological Implications for Capturing and Abstracting Human-AI Interaction Logs

This theoretical shift requires new methods for capturing and making sense of evidence. The new class of data —human-AI interaction traces—includes prompt sequences, the iterative refinement of prompts, the selection and revision of AI-generated outputs, and the full history of the interaction. Analyzing these traces is essential for assessing a new suite of emerging competencies central to modern knowledge work, such as prompt engineering, critical evaluation of AI outputs, and ethical integration of AI-generated content.

A central methodological challenge in working with any trace data is the "granularity dilemma": raw logs are often too fine-grained (e.g., individual mouse clicks) to be pedagogically meaningful, while overly coarse aggregations (e.g., total time spent) lose critical process details. The act of defining an "event" for analysis is thus an act of theoretical commitment. To address this in the new context of human-AI interaction, we propose four prescriptive rules for trace abstraction.

First, abstract to pedagogical moves. Low-level events should be aggregated into higher-level, semantically meaningful actions. For instance, a sequence of text-editing events followed by a "submit" click in a virtual lab could be abstracted into a single event: "Hypothesis Submitted".

Second, use time-windowing. To manage high-frequency data, events can be defined based on activity within discrete time windows, such as logging the primary activity type (e.g., "Reading," "Problem-Solving") within each 30-second interval.

Third, seed abstraction with human coding. For complex behaviors, human experts can manually code a subset of the data according to a predefined rubric. These labels can then be used as ground truth to train a machine learning classifier to automate the abstraction process for the entire dataset.

Fourth, abstract to human-AI interaction patterns. Specifically for generative AI traces, low-level logs of prompts and responses should be abstracted into meaningful interaction patterns. For example, a single, copied-and-pasted query could be labeled as a "Direct Prompting" event, whereas a sequence of queries that progressively refine the AI's output could be labeled as an "Iterative Refinement" event. Other patterns might include "Fact-Checking AI Output" (indicated by a student running a search query related to an AI's claim) or "Integrating AI Suggestion" (indicated by a student copying text from the AI and pasting it into their work document). This allows the assessment to focus on the quality of the student's engagement with the AI, not just the raw interaction log.

4 The Analytical Engine: A Methodological Toolkit for Educational Process Mining

Once digital traces have been collected and abstracted into meaningful event sequences, a range of AI methods can be applied to extract patterns and insights. The selection of a method must align with the specific assessment goal, the nature of the data, and a commitment to pedagogical utility.

4.1 A Principled Approach to Model Selection: Aligning AI Methods with Pedagogical Goals

In educational assessment, the reasoning behind a judgment is often as crucial as the judgment itself. This creates a fundamental tension between the predictive power of complex "black box" models and the interpretability of simpler ones. This trade-off underscores the critical importance of Explainable AI (XAI) in making powerful models transparent and actionable for educators and learners. An unexplainable prediction of "student at risk" is diagnostically useless; an explainable prediction that identifies the specific sequence of behaviors contributing to that risk is the foundation for effective intervention. Therefore, in this context, explainability is not merely a technical feature for building trust but a pedagogical necessity for enabling formative feedback. The following methodologies represent a toolkit for principled analysis, with a constant focus on the need for explainability.

Sequential Models (SPM, HMMs, RNNs/LSTMs): These models are ideal for understanding the temporal dynamics of learning. Sequential Pattern Mining (SPM) discovers frequently occurring ordered subsequences, which can identify common learning strategies or compare pathways between novice and expert performers. Hidden Markov Models (HMMs) infer a sequence of unobservable (hidden) cognitive or affective states (e.g., "Exploring," "Confused," "Planning") from observable student actions, allowing for the identification of unproductive learning loops. Recurrent Neural Networks (RNNs) and their variant, Long Short-Term Memory (LSTM) networks, are deep learning models that can capture complex, long-range dependencies in student interaction logs, making them highly effective for early prediction of students at risk of failure.

Clustering: Unsupervised clustering algorithms group students based on feature vectors derived from their process data (e.g., frequency of actions, time on task) without predefined labels. This data-driven approach can discover emergent learner profiles or problem-solving archetypes, such as "Systematic Planners" versus "Trial-and-Error Tinkerers," which can inform differentiated instruction.

Natural Language Processing (NLP): NLP techniques are essential for analyzing discourse traces from chat logs, written explanations, or think-aloud protocols. Sophisticated models like BERT can be trained on human-coded data to automatically classify student utterances according to a theoretical framework, enabling a scalable, nuanced assessment of collaboration quality or reasoning processes that goes far beyond simple participation metrics.

Multimodal Learning Analytics (MMLA): MMLA integrates data from multiple channels (e.g., video, eye-tracking, physiological sensors) with interaction logs to provide a holistic, synthesized assessment of constructs like engagement, frustration, and cognitive load.

4.2 From Black Box to Glass Box: The Pedagogical Requirement for Explainable AI (XAI)

Solving the "black box" problem is essential for building trust and accountability in educational AI. Explainable AI (XAI) techniques aim to make AI decisions understandable to humans, a feature that is paramount in educational contexts. For a teacher to act on an RNN's prediction that a student is "at-risk," they need to know why. XAI methods like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can analyze a model's decision and highlight which specific events in a student's recent history contributed most significantly to the high-risk prediction. Similarly, for NLP, attention visualization can highlight the specific words or phrases in a text that a model focused on to make its classification. This transforms an opaque prediction into a transparent, actionable diagnostic insight, turning the AI from an authority into a collaborative partner in the pedagogical process. Table 1 provides a comparative overview of these methodologies, emphasizing their application in CPS assessment and the critical need for explainability.

5 From Analytics to Action: Human-AI Orchestration in the Classroom

The ultimate value of AI-enabled PBA lies not in the sophistication of its analytics, but in its ability to positively impact teaching and learning. This requires translating analytical insights into concrete actions through well-designed feedback loops, adaptive systems, and a clear conception of the collaborative roles of teachers and AI. The effective implementation of these systems, however, hinges on developing a new form of professional capacity among educators—an "AI literacy" focused not on coding, but on data interpretation, probabilistic reasoning, and pedagogical translation.

Table 1: A Methodological Toolkit for AI-Enabled Process-Based Assessment

AI Methodology	Brief Description	Typical Digital Traces	Application in CPS Assessment	Key Advantages	Limitations & Critical XAI Requirements
Sequential Pattern Mining (SPM)	Identifies frequently occurring ordered subsequences of events.	Interaction logs (clickstreams, action sequences)	Discovering common learning strategies; comparing behavioral patterns across groups (e.g., novice vs. expert).	Inherently interpretable; captures temporality.	Can generate an overwhelming number of patterns; may ignore the relative frequency of different patterns.
Hidden Markov Models (HMM)	Infers a sequence of unobservable hidden states (e.g., cognitive states) from observable event sequences.	Interaction logs, coded behavioral sequences	Modeling student engagement, confusion, or planning phases; identifying unproductive learning loops.	Models latent cognitive states; captures dynamic processes.	State definitions require theoretical grounding; the Markov assumption can be limiting. XAI requires visualizing state transition diagrams.
RNNs/LSTMs	Deep learning models that process sequential data, capturing complex, long-term dependencies.	Time-series data from interaction logs; text sequences.	Early prediction of student success or failure; modeling the complex evolution of behavior over a whole course.	Highly effective at handling complex, nonlinear sequences; learns features automatically.	Prone to being a "black box." XAI methods like SHAP or LIME are nonnegotiable to identify which specific student actions contributed to a prediction, making the output actionable for teachers.
Clustering Algorithms	Groups data points based on similarity to identify natural clusters without predefined labels.	Feature vectors derived from logs, text, or multimodal data.	Identifying different problem-solving archetypes (e.g., "planners" vs. "tinkerers"); student segmentation for differentiated instruction.	Enables unsupervised discovery of emergent patterns.	Choice of cluster number can be subjective; results require domain expertise for interpretation. XAI involves characterizing clusters by their feature centroids.
Natural Language Processing (NLP)	Techniques enabling computers to understand, interpret, and generate human language.	Think-aloud protocols, chat logs, forum posts, written explanations.	Assessing collaboration quality; analyzing reasoning processes; providing feedback on argumentation.	Unlocks rich insights from qualitative data at scale.	Language is ambiguous and context-dependent. XAI can be achieved through attention visualization, highlighting key words/phrases that drove a classification.
Multimodal Learning Analytics (MMLA)	Integrates data from multiple channels (visual, auditory, physiological, logs) for a holistic view of learning.	Video, audio, eye-tracking, physiological signals, interaction logs.	Synthesized assessment of engagement, frustration, cognitive load, and collaborative synchrony.	Provides a richer, more nuanced learning profile; can improve predictive accuracy.	Data synchronization and fusion are complex; high analytical difficulty; significant privacy risks. XAI is critical to show the contribution of each modality to an inference.

5.1 The Teacher as Classroom Orchestrator: A New Model for Professional Practice

Effective implementation of AI-PBA necessitates a clear division of labor, shifting the teacher's role from being a manual data processor to that of a "classroom orchestrator" who leverages AI-driven insights to make informed pedagogical decisions. This model positions the teacher not as a passive recipient of AI outputs, but as an active director of a complex system of human and AI agents, preserving human pedagogical judgment as the central, authoritative element in the classroom. This new role requires a specific skill set: the ability to interpret AI outputs, understand their limitations and probabilistic nature, and translate data insights into pedagogically sound actions. This reconceptualization of the teacher's role represents a new theory of professional agency in AI-integrated classrooms, defining a complementary, rather than competitive, relationship between human and artificial intelligence.

5.2 Protocols for Human–AI Adjudication and Escalation

A proposed orchestration protocol, drawing on established models of teacher-AI collaboration, includes a clear division of labor and rules for interaction.

What the AI Decides: The AI system is best suited for automated, high-frequency, low-stakes tasks. This includes providing real-time, targeted hints based on a student's immediate actions ("One Teach, One Assist" model), offering path coaching by comparing a student's strategy to common successful patterns discovered through SPM, and flagging potential at-risk students based on predefined probability thresholds from an RNN model ("One Teach, One Observe" model).

What the Teacher Adjudicates: The human educator retains authority over high-stakes, context-dependent, and nuanced judgments. This includes interpreting complex or ambiguous patterns flagged by the AI, making final grading decisions, designing personalized, human-centered interventions for small groups ("Differentiated Teaching" model), and serving as the final arbiter in student appeals or disputes over an AI-generated assessment.

Escalation Rules: A clear protocol should govern the interaction between the AI and the teacher to ensure timely action without creating alert fatigue. For example: "If an RNN flags a student as 'at-risk' with a probability greater than 0.7 for two consecutive weeks, the system automatically adds an item to the teacher's dashboard prompting a one-on-one conference". This rule automates the initial detection but leaves the crucial interpersonal intervention to the human expert.

5.3 Engineering Explainable Feedback for Metacognitive Development

For feedback to be effective, it must be understandable and actionable. This requires translating the outputs of XAI tools into clear, narrative explanations for both students and teachers. Instead of presenting a complex technical plot from a SHAP analysis, the system should generate a plain-language summary.

Student-Facing Template: "The system suggested you might be stuck because your last five actions were all 'attempts' on the same sub-problem with no 'help-seeking' or 'example review' actions in between. Exploring a hint might reveal a new approach".

Teacher-Facing Template: "Student A was flagged as 'at-risk' this week. The key contributing factors from the model were a 50% decrease in time spent on planning activities compared to their baseline and a sequence of rapid, unsuccessful attempts on Problem 3, characteristic of a 'guessing' behavior pattern. A conversation focused on planning strategies may be beneficial".

By making the learning process visible and providing feedback on strategies rather than just outcomes, AI-PBA can be a powerful tool for fostering metacognition and a growth mindset. Traditional OBA can inadvertently reinforce a "fixed mindset" when students receive only a final score, which they may interpret as a judgment on their innate ability. In contrast, when students receive specific, process-oriented feedback, they come to understand that their methods and efforts can be analyzed and improved. This shifts their focus from "Am I smart?" to "What strategy can I try next?". AI-PBA tools that visualize progress over time and highlight specific improvements in strategy can be powerful facilitators of this crucial psychological shift, encouraging students to view challenges as learning opportunities rather than threats.

6 Building Trustworthy Systems: An Expanded Protocol for Validity, Governance, and Cultural Responsiveness

For AI-enabled PBA to be adopted responsibly and effectively, it must be built on a foundation of rigorous validation and a steadfast commitment to ethical principles. This section moves beyond broad statements to propose an operational protocol for ensuring the trustworthiness of these systems. The pursuit of "fair" and "unbiased" AI in education is often narrowly defined as achieving statistical parity across demographic groups within a single, usually Western, cultural context. This overlooks a deeper, more insidious bias: the encoding of culturally specific pedagogical values into a model's core logic. A truly trustworthy AI-PBA system must therefore be designed not for universality, but for adaptability and cultural attunement.

6.1 A Multi-Faceted Protocol for Validity and the "Meta-Validation" Challenge

Ensuring that an AI-PBA system measures what it purports to measure and performs reliably is paramount. A multi-faceted validation protocol is essential.

Construct Validity: This addresses the question: "Are we truly measuring the intended psychological construct?" Evidence can be gathered by correlating AI-PBA outputs with scores from established psychometric instruments (e.g., self-regulation questionnaires) or through qualitative analysis of think-aloud protocols, which can be compared to the strategies identified by the AI model.

Criterion Validity: This addresses the question: "How well do the AI's assessments agree with an external benchmark?" The most common benchmark is the judgment of human experts. A panel of trained educators can rate student process data, and these ratings are then compared to the AI model's output using statistical measures of agreement.

Generalizability and Robustness: A valid model must generalize across different tasks, student populations, and time. This involves testing model performance on out-of-sample tasks and implementing protocols for ongoing monitoring and periodic retraining to mitigate "model drift," where performance degrades as new data no longer resembles the original training data.

This process, however, reveals a deeper methodological challenge: the "meta-validation problem." If the validation protocol relies on human experts as the "gold standard," it begs the question of who validates the validators and accounts for their potential inconsistencies or biases. This points toward a need for more sophisticated validation approaches, such as using multiple methods in triangulation and critically examining the reliability of our own validation benchmarks.

6.2 Beyond Statistical Parity: The Imperative for Cultural Responsiveness in Assessment

Statistical fairness is a necessary but insufficient condition for ethical assessment. True fairness requires cultural responsiveness. Most AI in Education (AIED) systems are created by and for developed-world contexts, with Western cultural and pedagogical perspectives embedded in their design. The very constructs being measured—what constitutes "good" collaboration, "effective" help-seeking, or "productive" struggle—can be culturally dependent. For example, a model trained to value direct, argumentative discourse as a sign of strong collaboration may unfairly penalize students from cultures that value consensus-building and indirect communication. This means an AI-PBA model could show perfect statistical fairness across racial or gender groups within one culture, yet still be deeply biased against students from another.

Addressing this requires moving beyond purely technical debiasing to embrace participatory design methodologies. Participatory design involves the active collaboration of stakeholders—including teachers, students, and community members—in the design process to ensure the resulting technology meets their needs and aligns with their values. Co-designing assessment constructs and models with local communities can help ensure that the system's values align with the cultural context in which it is deployed. Furthermore, developing algorithmic impact assessment tools that are explicitly responsive to diverse cultural values and local wisdom is a critical area for future work.

6.3 An Operational Governance Framework for Ethical AI-PBA

Ethical considerations are not an add-on but a foundational requirement for AI-PBA. To move from abstract principles to concrete practice, the operational compliance checklist in Table 2 can guide the design, deployment, and

governance of these systems. This framework draws on established models for AI governance in higher education, such as the GOVAIHEI model, which delineates key domains for oversight. It ensures that student privacy, fairness, and transparency are protected through auditable actions.

Table 1: Ethics & Governance Compliance Checklist for AI-Enabled PBA

Category	Checklist Item and Operational Requirement
Consent & Transparency	Provide clear, plain-language consent forms detailing data types collected, purpose, retention period, and opt-out procedures. Students and parents must give informed consent before data collection begins.
	Implement student- and teacher-facing "Model Cards" that explain in accessible terms what a model predicts, its intended use, its known limitations, and its performance on different subgroups.
PII Handling & Security	Establish and enforce a protocol for pseudonymization of personally identifiable information (PII) at the point of data ingestion to protect student privacy.
	Implement strict, role-based access controls to ensure that only authorized personnel can access the level of data necessary for their legitimate function.
Bias & Fairness Audits	Conduct regular (e.g., per-term) audits of model performance across legally protected and pedagogically relevant demographic subgroups (e.g., by race, gender, socioeconomic status, prior knowledge level).
	Utilize established group fairness metrics to detect algorithmic bias. Define and document acceptable thresholds for performance disparities and a remediation plan if these thresholds are exceeded.
Explainability & Accountability	Ensure all automated assessments or high-stakes alerts are accompanied by an XAI-generated rationale that is presented in an understandable format to the end-user (student or teacher).
	Establish a clear, accessible, and well-documented process for students and teachers to appeal or request a human review of an AI-generated assessment or recommendation.
Human Oversight	Implement a formal protocol for escalating high-risk flags or ambiguous model outputs to a human educator for review and final decision-making.
	Maintain an immutable log of all human overrides of AI recommendations. This log should be periodically reviewed to identify potential systemic issues with the AI model or the oversight process itself.
Cultural Responsiveness	Conduct a participatory design process with target user communities to define key learning constructs and validate that they are culturally appropriate before model development.
	Validate model performance and interpretability across different cultural and linguistic groups before large-scale deployment to ensure it does not unfairly penalize different communication or problem-solving styles.

7 Discussion and Conclusion: Charting the Future of Assessment

7.1 Synthesis: Assessment as a Continuous, Diagnostic, and Co-Constructed Process

This paper has argued that the convergence of pedagogical need, rich digital trace data, and sophisticated AI methods —a trend now critically accelerated by the rise of generative AI—has created a pivotal moment for educational assessment. The traditional, outcome-focused paradigm is no longer tenable for many tasks, making a shift to process-based assessment a matter of necessity. We have proposed a principled, five-stage socio-technical design framework to guide the development of AI-enabled PBA systems. By grounding the framework in socio-technical systems theory

and the theory of distributed cognition, we have provided a robust theoretical foundation for this new paradigm. By mapping specific AI methodologies to concrete intervention pathways and embedding protocols for validity and ethical governance—including the novel imperative of cultural responsiveness—directly into the design process, this framework provides an actionable blueprint for moving beyond the limitations of traditional evaluation.

7.2 Implications for Educational Practice, Policy, and Research

The implications of this shift are significant. For educational practice, AI-PBA offers teachers powerful diagnostic tools to understand how their students learn, enabling more personalized, formative, and effective support. However, this requires a new professional literacy focused on data interpretation and pedagogical translation. For educational policy, it necessitates the development of robust governance frameworks, like the one proposed here, to ensure that these powerful technologies are used in a manner that is fair, transparent, and equitable. For educational research, it opens up new avenues for investigating the dynamics of learning at an unprecedented scale and level of detail, allowing us to test learning theories in more authentic and complex environments, particularly through the lens of human-AI distributed cognitive systems.

7.3 Future Directions: The Co-Evolution of Learning, Assessment, and AI

Looking forward, several key challenges and opportunities will define the future of this field. A primary frontier is the assessment of human-AI collaboration itself. As students increasingly use generative AI as a partner, the focus of PBA must expand from assessing what a student knows to assessing how they build knowledge in partnership with an AI cognitive tool. Second, as the field matures, it must confront the "meta-validation" problem, developing more robust methods to ensure that the benchmarks we use to validate our systems are themselves valid and unbiased. Third, the field must move toward cross-cultural adaptation, designing systems that are not monolithic but are culturally aware and responsive to diverse learning contexts through participatory design.

Ultimately, the widespread adoption of robust AI-PBA could become a powerful driver for systemic curricular and pedagogical reform. There is a well-known phenomenon in education: assessment drives instruction. For decades, educators have advocated for teaching complex skills like iteration, strategic thinking, and collaboration, but the assessment system, optimized for scalable OBA, has incentivized the teaching of more easily measurable content knowledge. By creating a scalable and reliable method to measure the process of these complex skills, AI-PBA fundamentally alters the incentive structure of the educational system. It makes teaching these skills a priority because they can now be valued and assessed. In this way, AI-PBA can be a catalyst for fundamentally rethinking the purpose and practice of assessment, aligning the entire educational enterprise more closely with the development of the deep, transferable skills required in the 21st century. The path forward requires a symbiotic partnership between educators, researchers, and technologists, working collaboratively to build assessment systems that are not only more intelligent but also more equitable and humane.

To Cite This Article Alexander LEE. (2025). From Traces to Teaching: A Socio-Technical Framework for Process-Based Assessment in an Age of Distributed Cognition. *Artificial Intelligence Education Studies*, *1*(3), 43-55. https://doi.org/10.6914/aiese.010304

Reference

[1] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138 52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052 1092. https://doi.org/10.1007/s40593-021-00286-9

[3] Benjamin, R. (2020). *Race after technology: Abolitionist tools for the new Jim Code*. Polity Press.

[4] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77 91). PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html

[5] Dillenbourg, P., & Jermann, P. (2010). Technology for classroom orchestration. In S. R. F. a. M. J. Nathan (Ed.), *The new science of learning* (pp. 525 552). Springer.

[6] Dimitriadis, Y., et al. (2021). Orchestrating technology enhanced learning: A literature review and a conceptual framework. *International Journal of Computer-Supported Collaborative Learning, 16*(3), 307 342. https://doi.org/10.1007/s11412-021-09353-0

[7] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214 226). https://doi.org/10.1145/2090236.2090255

[8] Farazouli, A., et al. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Computers & Education: Artificial Intelligence,* 7, 100223. https://doi.org/10.1016/j.caeai.2024.100223

[9] Fischer, C., et al. (2020). Mining sequential patterns in educational data: A systematic review and future research agenda. *Journal of Educational Data Mining, 12*(1), 1 34. https://doi.org/10.5281/zenodo.3946221

[10] Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 225 234). https://doi.org/10.1145/3303772.3303791

[11] Holstein, K., & Aleven, V. (2022). Co-orchestrating the AI-enhanced classroom: A framework for teacher-AI collaboration. *Computers & Education, 185*, 104521. https://doi.org/10.1016/j.compedu.2022.104521

[12] Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1 14). https://doi.org/10.1145/3290605.3300327

[13] Hutchins, E. (1995). *Cognition in the wild*. MIT press.

[14] Järvelä, S., et al. (2023). Hybrid human-AI regulation in collaborative learning. *Educational Technology Research and Development,* 71(4), 1435 1453. https://doi.org/10.1007/s11423-023-10255-8

[15] Karumbaiah, S., & Brooks, C. (2021). The implications of algorithmic bias in education. In *Companion Proceedings of the 11th International Conference on Learning Analytics & Knowledge* (pp. 637-643). https://www.solaresearch.org/wp-content/uploads/2021/04/LAK21_Companion_Proceedings.pdf#page=650

[16] Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[17] Kim, J., & Lee, H. (2024). Types of Teacher-AI Collaboration in K-12 Classroom Instruction: A Qualitative Study. *Education and Information Technologies, 29*(2), 1547 1569. https://doi.org/10.1007/s10639-023-12093-4

[18] Kling, R. (1980). Social analyses of computing: Theoretical perspectives in recent empirical research. *ACM Computing Surveys, 12*(1), 61 110. https://doi.org/10.1145/356802.356806

[19] Lodge, J. M., & Corrin, L. (2017). What is the role of the teacher in a distributed learning environment? *Australasian Journal of Educational Technology, 33*(5). https://doi.org/10.14742/ajet.3541

[20] Matson, J. O. (2025). Where Meaning Emerges: Human AI Dialogue, Distributed Cognition, and the Cognitive Intraface. *Postdigital Science and Education*. https://doi.org/10.1007/s42438-024-00508-y

[21] Mehrabi, N., et al. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1 35. https://doi.org/10.1145/3457607

[22] Mitchell, S., et al. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141 163. https://doi.org/10.1146/annurev-statistics-042720-125902

[23] Molenaar, I., & Knoop-van Campen, C. (2018). How teachers use a learning analytics dashboard to support student' s self-regulated learning. In *Proceedings of the 8th International Conference on Learning Analytics & Knowledge* (pp. 33 42). https://doi.org/10.1145/3170358.3170425

[24] Mumford, E. (1983). *Designing human systems for new technology: The ETHICS method*. Manchester Business School.

[25] Nartey, E. K. (2025). *Generative AI in Higher Education: Guiding Principles for Teaching and Learning*. Routledge.

[26] Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[27] OECD. (2017). *PISA 2015 Results (Volume V): Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264285521-en

[28] Pasmore, W. A., Francis, C., Haldeman, J., & Shani, A. (1982). Sociotechnical systems: A North American reflection on empirical studies of the seventies. *Human Relations, 35*(12), 1179 1204. https://doi.org/10.1177/001872678203501207

[29] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135 1144). https://doi.org/10.1145/2939672.2939778

[30] Ropohl, G. (1999). Philosophy of socio-technical systems. *Techné: Research in Philosophy and Technology, 4*(3), 59 72. https://doi.org/10.5840/techne19994318

[31] Sarmiento, J. P., & Wise, A. F. (2022). A review of participatory design in learning analytics. *Journal of Learning Analytics, 9*(3), 534 551. https://doi.org/10.18608/jla.2022.7661
[32] Trist, E. L. (1981). The evolution of socio-technical systems. In A. H. Van de Ven & W. F. Joyce (Eds.), *Perspectives on organization design and behavior* (pp. 19 75). Wiley.
[33] UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000386693
[34] Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277 304). Lawrence Erlbaum Associates Publishers.
[35] Wise, A. F. (2014). Designing pedagogical interventions to support student use of learning analytics. In *Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 203 211). https://doi.org/10.1145/2567574.2567588
[36] Zhai, X. (2023). ChatGPT for next generation science learning. *XR and AI in Education, 1*(1), 1-13. https://doi.org/10.1007/s44279-023-00001-3
[37] Zhang, J., & Patel, V. L. (2006). Distributed cognition, representation, and affordance. *Pragmatics & Cognition, 14*(2), 333 341. https://doi.org/10.1075/pc.14.2.11zha

Editor *Changkui LI wtocom@gmail.com*