

Human–AI Learning Assessment

Testing When Evidence from Student–AI Dialogue Can Support Defensible Proficiency Inferences

Proposer: Fabrizio Terzi, independent researcher, Pyragogy **Contact:** info@pyragogy.org

Repository: <https://github.com/pyragogy/human-ai-learning-assessment> **Date:** October 2026

Proposed solution and PISA relevance

HALA proposes a bounded OECD co-design programme to establish when evidence from student–AI dialogue can validly support PISA proficiency inferences. Existing assessment frameworks—PISA 2015 Collaborative Problem Solving (controlled computer agents), dynamic assessment (scaffolded performance as evidence), PISA 2025 Learning in the Digital World (process and response evidence)—show that interaction data can support proficiency claims under specific design conditions.¹ Building on Evidence-Centred Design and argument-based validation, HALA tests a narrower question: can *generated, adaptive AI dialogue*—where the assistant proposes explanations, counterexamples and requests—yield evidence that supports valid inference about *learner proficiency*, and under what design conditions?

Two candidate constructs. The programme evaluates both **unaided proficiency** (performance on independent probes without AI assistance) and **AI-supported problem-solving proficiency** (evidence evaluation and problem solving under specified AI support). Which observations warrant which inference is the object of study, jointly specified with the OECD before any data collection. Agent variance is not assumed to be a contaminant: scaffold responsiveness may itself be construct-relevant. The original T2 learning study remains separate.

Experimental skeleton. A proposed design crosses three task families (science source appraisal, correlational reasoning, controlled comparison) with agent policy (explanatory assistant, critical collaborator, generic management) and assistance level; language versions (English and an agreed partner language) require adaptation. Analyse turns clustered within students and student/task/rater/agent variance; log actual agent outputs and model/prompt versions. Counterbalanced unaided probes provide validity anchors with order/reactivity checks. Calibrate raters; mask condition metadata where feasible while preserving support context. Preregister reliability, agent-task effects, anchor relations and differential language functioning, with construct-specific retain/revise/reject rules, uncertainty criteria and sample/precision rationale agreed before pilot collection; no numerical thresholds or sample sizes are invented.

Observation and provenance. The proposed manual annotates each student turn with a behaviour code, the preceding AI intervention, and the flag `learner_origin = indeterminate`—the system does not identify cognitive authorship. A separate field `code_status` distinguishes `sufficient_evidence` (code fires with adequate behavioural basis) from `insufficient_evidence` (short assent, off-task, or response too minimal to evaluate). Draft codes include **EVIDENCE_REQUEST** (student requests grounds), **COMPETING_CAUSE** (proposes rival explanation), **DISCRIMINATING_TEST** (designs a comparison), **SOURCE_APPRAISAL** (evaluates credibility), **JUSTIFIED_REVISION** (revises with evidence), and **CALIBRATED_UNCERTAINTY** (proportional confidence). These are candidate codes, not validated scores; the existing heuristic fixture is a falsification instrument.

Worked synthetic task and rubric. Fictional cards report a screen-use/sleep correlation and no clear sleep difference after randomly assigned screen restriction. The student evaluates the causal claim. AI: “Could a third factor, such as evening use, explain the correlation?” Student: “Maybe kids using screens at night also have routines that leave less time for sleep.” Tag **COMPETING_CAUSE** for a plausible rival explanation addressing the same pattern without unsupported certainty; record AI-suggested support and `learner_origin=indeterminate`. The evening/night echo suggests possible uptake, not copying or cognitive authorship. “Yeah, maybe” receives `code_status=insufficient_evidence`, not a zero ability score.

Measurement development. Codes are developed through a predeclared five-step sequence: (1) define minimum observation opportunities per code; (2) annotate student action and preceding AI intervention separately; (3) estimate inter-rater agreement and code-distribution shifts under different agents and prompt versions; (4) compare conversational indicators against independent unaided probes and examine language comparability; (5) retain only indicators supporting the intended construct. Unsupported unaided interpretations are rejected; replicable responsiveness may support the assisted construct. No universal agent-variance cutoff is assumed. Negative results are reported explicitly.

¹OECD (2017), *PISA 2015 Assessment and Analytical Framework*, [doi:10.1787/9789264281820-en](https://doi.org/10.1787/9789264281820-en); Poehner & Lantolf (2005), *Dynamic assessment in the language classroom*, [doi:10.1191/1362168805lr1660a](https://doi.org/10.1191/1362168805lr1660a); OECD (2026), *PISA 2025 Assessment and Analytical Framework*, [doi:10.1787/86c36975-en](https://doi.org/10.1787/86c36975-en).

Collaboration with OECD

Three bounded work packages are proposed:

Package 1 — Construct and estimand agreement. OECD specifies target competency constructs, delivery requirements, multilingual equivalence standards, and safeguarding expectations. HALA produces a measurement-design document distinguishing unaided and AI-supported constructs and defining the candidate `learner_origin` and fallback rules.

Package 2 — Tasks, prompts, and observation instrument. HALA delivers three task specifications with version-controlled prompt registries, a behavioural coding manual with worked examples, a reproducible synthetic trace pipeline, and an adversarial test suite for testing AI-supplied evidence—all developed through iterative co-design with OECD experts. OECD reviews task constructs, prompt policies for cross-language consistency, and data-security requirements.

Package 3 — Instrument review and pilot-readiness gate. HALA supplies a rater-calibration and reliability-analysis plan; an empirical report requires approved data and specialist capacity. It also supplies a data-protection design (DPIA-ready framework, pseudonymisation, lawful-basis analysis, controller/processor allocation, contractual training exclusion, retention/deletion schedules and transfer controls), and a written assessment of conditions under which a student-facing pilot becomes feasible. Authorised OECD/educational partners would need to provide pilot access, psychometric expertise and ethics processes; none is secured. **No pilot deployment can proceed without agreed resources, partners, ethics approvals, and financing. Timing and workload for Packages 1–2 depend on agreed capacity; no staffed team or service is promised. Cooperation is unpaid.**

Intellectual property and model choice

IP: Source code and documentation developed before collaboration (background) remain private without an open-source licence; ownership and a bounded, non-exclusive OECD research licence for co-developed designs (foreground) require written agreement before joint work; no rights are granted here. Third-party model rights and provider terms are not licensed by the proposer.

Models: No model or provider is selected. HALA prioritises EU-hosted and independently deployable candidates in a provider-independent strategy. Mistral Large 4 (released 6 October 2026, public preview, API alias `mistral-large-4`; weights and self-deployment rights must be rechecked) is a candidate, not a selection. Other candidates include OpenAI GPT-4o snapshot `gpt-4o-2024-08-06` and at least one independently deployable open-weight model; none is preselected. Model versions and interaction policies are fixed within each evaluative phase; updates require a documented new phase. EU data residency and training exclusion must be verified, not assumed from provider nationality.

Full-scale cost dimensions: model inference tokens (including dialogue history, retries and auxiliary calls), human coding and rater moderation, translation and adaptation, psychometric analysis, platform operations and monitoring, and security review. No per-session price or zero-cost self-hosting is presumed. A detailed cost formula for reproducible inference budgets is maintained in the project assumptions document; published tariffs, concurrency, latency and synthetic workloads inform a full budget before deployment. Provider withdrawal or forced updates pause assessment; substitutes require renewed comparability tests.

Privacy, security, and feasibility

Current state: synthetic prototype only—private VPS, no real participant data, no personal information stored. Architecture: two-process separation with HMAC-authenticated IPC and internal security review; no independent certification is claimed. OS-level user isolation is not deployed.

Preconditions for any student-facing pilot: completed DPIA, pseudonymisation, lawful basis under GDPR/DPA with controller/processor allocation (not automatic parental-consent assumption), contractual and technical exclusion of student-data training (including feedback/sharing), retention/deletion schedules, subprocessor and international-transfer controls, incident-response procedures, child-safeguarding mechanisms (prompt monitoring, escalation, access logging, human-review boundaries), and appropriate ethics approvals and institutional partnerships. Pseudonymised dialogue remains potentially personal data. Providers receive only approved minimal context, never identity keys or withheld unaided items. These safeguards are pre-pilot gates. The prototype is a research instrument for a prospective programme, not an operational assessment platform.

Readiness. This proposal describes a planned research programme, not an operational system. No validated scores, student data, ethics approval, institutional partnership, or deployed AI tutor exist. The research question—whether evidence from AI-mediated dialogue can validly support proficiency inferences—is the object of study, not a capability claim. Controlled repository access can be offered to OECD evaluators on request.