

# **Governance-Integrated Artificial Intelligence for Assessing Transdisciplinary Competencies in Project-Based Learning: A Design Science and Pilot Validation Study**

Dr. Ayman Abu Hajaj

Affiliation: Ono Academic College, Kiryat Ono, Israel

Email: ayman2424@gmail.com

ORCID: 0009-0002-1944-8473

Role: First Author and Corresponding Author

Nadine Sheikh

Affiliation: Girne American University, Kyrenia, Cyprus

Email: nadinedia@gmail.com

ORCID: 0009-0006-1527-7514

Role: Co-author

## **Corresponding Author:**

Dr. Ayman Abu Hajaj

Ono Academic College, Kiryat Ono, Israel

Email: ayman2424@gmail.com

## **Abstract**

Assessing transdisciplinary competencies in project-based learning (PBL) is difficult because critical thinking, collaboration, and creativity develop across time, artefacts, and social interaction rather than in a single test event. Artificial intelligence (AI) can integrate longitudinal evidence, but consequential use raises concerns about construct validity, proxy bias, transparency, privacy, and institutional accountability. This design science study develops a governance-integrated assessment architecture that couples construct-aligned evidence, provisional score generation, fairness and explanation checks, and authorised human review within one documented cycle. Six design requirements were refined through a three-round Delphi process with 12 experts. A newly specified agent-based simulation involving 500 synthetic learners, five assessment cycles, and 50 replications compared the full architecture with a matched no-governance condition and three governance ablations under varied task-complexity, collaboration, and stress conditions. In the full-governance arm, predictive agreement with synthetic latent competency ranged from  $r = .934$  to  $.935$  across competencies, mean absolute error was  $.059$ , temporal fidelity ranged from  $.918$  to

.919, decision coverage was .820, and every numeric decision had an explanation record. Relative to no governance, agreement increased by .025-.026 and mean absolute error decreased by .011-.012, but capacity-constrained review reduced decision coverage by .173. Fairness-composite differences were small and their Monte Carlo intervals included zero. A field pilot with 47 undergraduates in three interdisciplinary PBL courses showed substantial correlations with independent instructor ratings ( $r = .71-.79$ ); reported McDonald's omega coefficients ranged from .76 to .81. The findings support operational feasibility and an initial validity argument, not real-world effectiveness, protected-group fairness, or automated high-stakes use.

Keywords: artificial intelligence in education; project-based learning; transdisciplinary competencies; learning analytics; construct validity; algorithmic fairness; human governance

## 1. Introduction

The demands on higher education increasingly call for solving problems that span across disciplines, engaging in integrating non-uniform evidence, sustained collaboration, and delivering unique and defensible solutions. Project based learning (PBL) is well aligned with these goals since learners experience extended periods of inquiry, design, communication, feedback, and revision (Kokotsaki et al., 2016; Perignat & Katz-Buonincontro, 2019; Rababah, 2022; Rababah et al., 2023). Instead, however, often assessment is focused on the final product or a few observations by the instructor. These snapshots can make it difficult to see how critical thinking, collaboration and creativity are manifested during development.

Digital learning spaces provide an expanded trail of evidence. Collaboration spaces can record who did what and how it was reciprocated; research logs can show the scope and quality of the sources used; and version histories can show substantive changes. Learning analytics can therefore more clearly be used to make longitudinal processes more visible (Gašević et al., 2015; Ifenthaler & Yau, 2020; Rababah, 2023; Rababah & Rababah, 2024; Viberg et al., 2018). However, the plethora of data doesn't solve the measurement problem. A trace is not a competency, a prediction is not a valid score interpretation and a technically accurate estimate is not necessarily a valid educational decision. The meaning and effects of using AI for assessment can be distorted, due to misaligned proxies, historical inequalities, absence of face-to-face activity or overreliance on institutions (Baker & Hawn, 2022; Jacobs & Wallach, 2021; Suresh & Gutttag, 2021).

The central question is thus how to implement a sociotechnical challenge. The hallmark of responsible assessment is the explicit construct model, documented data provenance, appropriate uncertainty, subgroup monitoring, explanations that educators and learners can use, and an allocation of authority that allows for real corrections and appeals. Human-centred AI research also focuses on the importance of embedding stakeholder involvement, human control, safety, and trustworthiness in the system life cycle, as opposed to being added on in the wake of deployment (Alfredo et al., 2024; Holmes et al., 2022). The point of end-to-end auditing and model documentation is that governance is a chain of operational controls, records and decisions, and not a statement of ethical principles (Mitchell et al., 2019, Raji et al., 2020).

This study responds by developing a governance-integrated AI architecture for assessing three enabling transdisciplinary competencies - critical thinking, collaboration, and creativity - in interdisciplinary PBL. Transdisciplinary does not mean that these three competencies cover all of the domain. Instead, it refers to capacities that students will develop that cut across disciplines—capacities that will enable students to connect knowledge and action across disciplines. Architecture is the provision of the algorithmic outputs as provisional evidence. The evidence selection, score generation, fairness review, explanation, instructor judgement, override and contestation are interconnected and mutually updated and maintain a shared audit trail. This controlled coupling is an extension of the traditional pipeline model: If evidence quality changes, or if the explanation is implausible, or if a human override is necessary, then there should be a way to review it again, not forgotten after a score has been produced.

The approach of design science research was chosen due to the goal of creating and testing an artefact to solve a practical and theoretical problem (Hevner et al., 2004; Peffers et al., 2007). The three components of the evaluation were: refinement of the design requirements and indicator weights by experts, a small-scale simulation of the coherence of the indicators during the operation, and a pilot in the field of initial measurement evidence. The pilot was not designed as a test of intervention effectiveness as it was single-site and under powered for confirmatory group comparisons.

The study addresses three research questions:

1. RQ1. Which design requirements are necessary for a governance-integrated system that uses longitudinal evidence to assess critical thinking, collaboration, and creativity in interdisciplinary PBL?
2. RQ2. Can the proposed architecture execute a complete cycle of score generation, governance checks, explanation, and review in a controlled agent-based proof-of-concept environment?
3. RQ3. Do the resulting competency scores show preliminary convergence with independent instructor ratings and sufficient descriptive coherence to justify a larger validation study?

## 2. Conceptual and Empirical Foundation

### 2.1 Competency Enactment in Project-Based Learning

Interdisciplinary learning brings together the lenses of established disciplines, while transdisciplinary learning also redefines the problem from a systems, stakeholders and action perspective that is not amenable to any single discipline. In this kind of work, critical thinking, collaboration and creativity are facilitatory skills. Critical thinking helps in assessing and synthesizing evidence, collaboration disseminates knowledge and facilitates interdependent work, and creativity produces and crafts alternatives that are novel and relevant. The capacities are repeatedly mentioned in international competency frameworks and this is not enough to define an assessment model (Voogt & Roblin, 2012).

PBL emphasizes temporal evidence, particularly. The finished product might be good, but it could be plagiarized, full of logic gaps, or heavily edited by the teacher. On the other hand, careful research, fair collaboration, and thorough rewriting can produce a small product. This blind spot can be mitigated with evidence from the longitudinal data, provided that the chain of evidence from construct to indicator to score interpretation is clearly defined. The Standards for Educational and Psychological Testing place their emphasis on the use and interpretation of scores and not on the instrument itself (American Educational Research Association et al., 2014). The framework, therefore, does not simply pose the question "does a digital signal predict a rating," but "what does a digital signal predict, what does it not predict, and, is the proposed use justifiable?"

These indicators are not regarded as isolated effects of a single latent variable, but rather as formative components of a composite, as a set of indicators that jointly depict the assessed representation of CT, rather than as separate indicators. The same is true for the composites of collaboration and creativity. This is important because internal-consistency coefficients do not constitute a perfect measure of reliability for formative composites (Bollen & Lennox, 1991). Rather, evidence needs to be obtained via content justification, stability, relations with external criteria, sensitivity analysis, and consequences of use.

AI-powered evaluation could handle more diverse evidence streams that are not the kind that can be continually monitored by an instructor within a class cohort. Tree based ensemble methods are potential architecture for tabular indicators given their ability to capture nonlinearities and interactions, and the ability to use feature-attribution methods. However, predictive fit is just one component of the validity argument. A model can simulate instructor ratings while using proxies that are not relevant to the competency such as opportunity to learn, platform access, writing fluency, etc. Measurement and fairness are thus linked: construct-irrelevant variance and construct underrepresentation can be harms that are systemic at the group level when access to observable evidence varies systematically (Jacobs & Wallach, 2021). Explanatory needs should also be audience- and purpose-specific. Global explanations can be used to help carry out a model audit

by highlighting features which tend to affect scores, while local explanations can be used to help review a specific learner estimate (Lundberg & Lee, 2017; Ribeiro et al., 2016). Technical fidelity is required but not enough. In education, an explanation needs to be linked to evidence that is to be recognized by a project, admit uncertainty and limitations, and provide an opportunity to correct an error (Khosravi et al., 2022). Explanation coverage thus does not provide evidence that the explanation is understandable or pedagogically appropriate. The current artefact is an abstraction of the general architecture from the field-pilot instantiation. The candidate models are Random Forest and Gradient Boosting (GB), when the training data is large enough, partitioning is possible at the learner level. This pilot sample of 47 is too small to fit and validate such models. Thus, scores from the field were normalised and assigned an average Delphi weighting, and the simulation focused on whether the overall architecture could perform the proposed sequence of computational- and governance-related tasks. This separation is to prevent the small pilot study from being presented as a machine-learning validation study. Fairness, Human Oversight and Institutional Governance (2.3) Algorithmic fairness is not one size fits all. These are various normative commitments that can be incompatible with one another in the presence of base rate differences (Chouldechova, 2017; Dwork et al., 2012; Hardt et al., 2016; Kleinberg et al., 2017; Mitchell et al., 2021). Only when the criteria, reference groups, thresholds, and educational consequences are documented, can a fairness index be interpreted. There is no justification for a mitigation to be accepted just due to the fact that it makes a statistic better; reweighing can help resolve representation imbalance without relabelling the observations (Kamiran & Calders, 2012). It also needs to maintain the meaning of the construction and not mask what is a normal uncertainty. Human monitoring is also more onerous than having a teacher at the end of the pipeline. Oversight can be "ceremonial" when reviewers don't have the time, evidence, power to deviate from the system, or institutional backing to disagree (Green, 2022). Access to role, documented reasons for access or override, separation of development and approval in appropriate cases, and a pathway for learners who are affected by the decision to seek explanation, data corrections, reconsideration. These are conditions that align with lifecycle risk management and guidance from the international community, and with the recognition of some uses of AI in the field of educational evaluation as high risk (European Parliament & Council of the European Union, 2024; Tabassi, 2023; UNESCO, 2023). Three principles of sociotechnical design of decision systems are thus adopted for the framework. Three principles of sociotechnical design of decision systems are therefore adopted for the framework. First, it is important to realize that the coupling is two-way: evidence rules, thresholds and features may be updated based on an analysis of human review and/or data-quality findings. Secondly, there's documentation of the versions of models, policies, data, and rubric, and changes must be authorized. Third, the system can choose to abstain, suspend automated processing, or refer a case for further review if the evidence is not sufficient or the governance threshold is exceeded. These principles avoid governance from being a post hoc ethics audit and ensure there is no tight integration that leads to the spread of an undetected error throughout the entire workflow. 2.4 Positioning and Research Gap Despite the rapid growth of research on the use of AI in higher education, there are continued concerns around

limited educator involvement, variable ethical assessment, and a lack of external validation (Bond et al., 2024; Zawacki-Richter et al., 2019). Learning analytics provides longitudinal evidence; fairness research provides formal criteria; explainable AI provides interpretive tools; and governance frameworks provide lifecycle controls. PBL assessment provides relevant tasks and rubrics. The un-resolved design problem is how these contributions should work as a singular reviewable measurement process. The gap addressed with the artefact is defined in Table 1.

**Table 1**

*Conceptual Positioning of the Governance-Integrated Assessment Architecture*

| <b>Tradition</b>     | <b>Principal contribution</b>                        | <b>Unresolved problem in PBL assessment</b>                             | <b>Design response</b>                                                  |
|----------------------|------------------------------------------------------|-------------------------------------------------------------------------|-------------------------------------------------------------------------|
| AI in education      | Prediction, classification, and feedback             | Model performance may be evaluated apart from construct meaning and use | AI scores remain provisional evidence within a validity argument        |
| Learning analytics   | Longitudinal, multi-source process evidence          | Behavioural traces can become unjustified proxies                       | Every feature is linked to a construct, evidence source, and limitation |
| Algorithmic fairness | Formal disparity criteria and mitigation             | Criteria may conflict or lack educational justification                 | Metrics, groups, thresholds, and trade-offs must be documented          |
| Explainable AI       | Global and local interpretation                      | Explanations may be technically faithful but unusable                   | Explanations are tied to project evidence, review, and contestation     |
| AI governance        | Lifecycle risk management and accountability         | Principles may remain detached from assessment operations               | Versioning, audit logging, authority, override, and appeal are embedded |
| PBL assessment       | Authentic tasks, rubrics, and professional judgement | Continuous process evidence is difficult to review at scale             | AI organises evidence without replacing educator judgement              |

### 3. Design Science Development of the Artefact

**3.1 Design Science Logic** The study process was carried out in accordance with the design science process of problem identification, problem definition, design and development, demonstration, evaluation and communication (Hevner, 2004; Peffers, 2007). The artefact is a trained model, which is not a single model. It's an assessed workflow with a competency model, evidence rules, provisional scoring, predictive models for subsequent scaling, fairness and explanation controls, human-review protocol and an audit trail. This definition sets the standard for assessment—the whole assessment process instead of accuracy alone. A set of six design requirements came from the educational measurement, learning analytics, fairness, explainability and governance scholarship. Six design requirements were identified from the educational measurement, learning analytics, fairness, explainability and governance scholarship and further elaborated with expert consultation. Table 2 provides a mapping of each requirement to an operational mechanism and an evaluation question. In design science, the traceability is important because the artefact has to demonstrate the output of the design process as well as the mechanisms used to solve the problem for which it was designed.

*Table 2*

*Design Requirements, Operational Mechanisms, and Evaluation Logic*

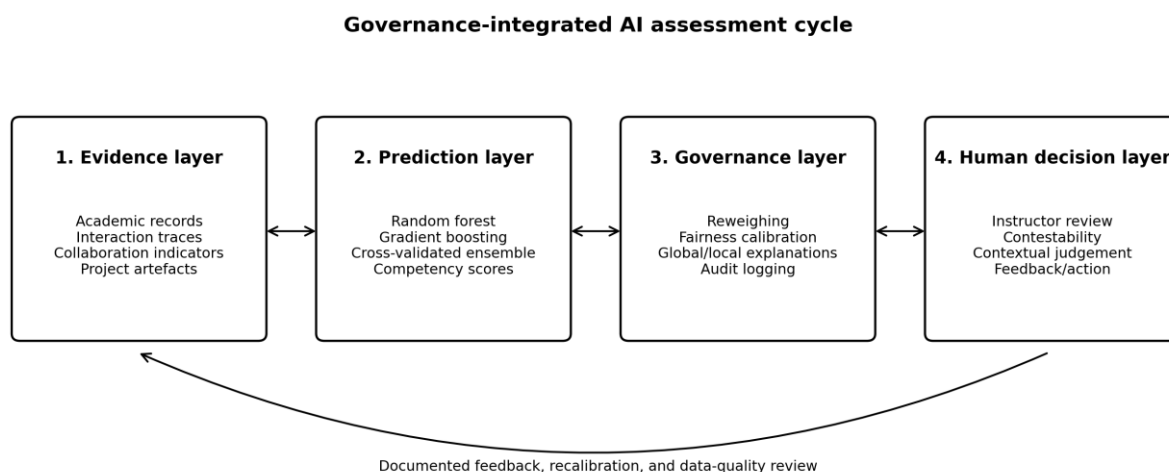
| Requirement                  | Operational mechanism                                               | Primary risk addressed                            | Evaluation evidence                                                                                               |
|------------------------------|---------------------------------------------------------------------|---------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| DR1 Longitudinal evidence    | Evidence collected across the project cycle                         | Snapshot bias and hidden process variation        | Complete multi-cycle processing in simulation; 10-week collection in pilot                                        |
| DR2 Construct alignment      | Feature-to-construct specification and human challenge              | Invalid proxies and construct underrepresentation | Expert review, indicator dictionary, external instructor ratings                                                  |
| DR3 Fairness-aware modelling | Reweighting, prespecified disparity monitoring, bounded calibration | Unequal error or opportunity                      | Prespecified component gaps, matched comparison, and fairness-monitor ablation; real subgroup validation deferred |
| DR4 Explanation              | Global and learner-level feature attribution                        | Opaque or unreviewable output                     | Explanation-record coverage and no-explanation ablation; educator-facing reports in pilot                         |

| Requirement                            | Operational mechanism                               | Primary risk addressed                            | Evaluation evidence                                                                      |
|----------------------------------------|-----------------------------------------------------|---------------------------------------------------|------------------------------------------------------------------------------------------|
| DR5 Human authority and contestability | Confirm, amend, reject, abstain, and appeal routes  | Automation bias and accountability gaps           | Capacity-constrained review, abstention, and no-structured-review ablation               |
| DR6 Auditability and data governance   | Data minimisation, versioning, provenance, and logs | Irreproducibility, privacy risk, and silent drift | Code, parameters, exact seeds, row-level synthetic outputs, hashes, and run log archived |

### 3.2 Coupled Architecture

The architecture contains four coupled functions. The evidence function acquires authorised academic records, project artefacts, interaction traces, and collaboration indicators, while minimising collection and retaining raw communication content only when no less intrusive feature is adequate. The scoring and inference function normalises indicators, computes provisional competency composites, and, in a mature implementation, may apply learner-level partitioned ensemble models. The assurance function conducts data-quality checks, fairness monitoring, bounded calibration, explanation, and version logging. The human decision function enables an authorised instructor to confirm, amend, reject, or defer the estimate and provides a documented route for clarification and appeal.

The arrows among functions are reciprocal rather than decorative. An implausible explanation can initiate evidence review; an override can reveal a construct or data-quality problem; and a change in policy or rubric can invalidate a feature definition. Every final score must therefore be linked to the evidence window, transformation rules, scoring version, fairness checks, explanation, reviewer, and disposition.

**Figure 1***Governance-Integrated AI Assessment Cycle*

**Note.** Bidirectional arrows represent controlled feedback. Human review, recalibration, and data-quality findings can return the process to an earlier function; automation may be suspended when evidence or governance thresholds are not met.

### 3.3 Competency Operationalisation

Each competency was represented by three indicators and a Delphi-derived weight. Features were normalised to [0, 1] within the applicable reference sample before aggregation. These weights are context-specific design parameters, not universal estimates. The score formulas are  $CT = .35(\text{Inquiry depth}) + .35(\text{Source diversity}) + .30(\text{Substantive revision})$ ;  $COL = .40(\text{Task balance}) + .35(\text{Interaction frequency}) + .25(\text{Response timeliness})$ ; and  $CRE = .45(\text{Novelty rate}) + .30(\text{Solution diversity}) + .25(\text{Lexical originality})$ . Table 3 clarifies both intended meaning and principal validity threat.

**Table 3***Provisional Competency Indicators and Delphi Weights*

| Competency        | Indicator (weight)     | Operational meaning and evidence                                                    | Principal validity threat                                      |
|-------------------|------------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------------|
| Critical thinking | Inquiry depth (.35)    | Depth and specificity of task-relevant questions in discussion and revision records | Verbosity or instructor prompting mistaken for reasoning depth |
| Critical thinking | Source diversity (.35) | Breadth and balance of credible source categories in citations and research logs    | Quantity or access mistaken for evidential quality             |

| Competency        | Indicator (weight)          | Operational meaning and evidence                                                     | Principal validity threat                                    |
|-------------------|-----------------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------|
| Critical thinking | Substantive revision (.30)  | Revisions that alter reasoning, evidence, or solution structure in version histories | Revision count without attention to quality or need          |
| Collaboration     | Task balance (.40)          | Distribution of contributions without excessive concentration in task logs           | Equal quantity mistaken for equitable or expert contribution |
| Collaboration     | Interaction frequency (.35) | Task-relevant reciprocal exchanges in team communication                             | High message volume mistaken for effective coordination      |
| Collaboration     | Response timeliness (.25)   | Promptness of response to team requests with outlier control                         | Time-zone, disability, care, or access differences           |
| Creativity        | Novelty rate (.45)          | Non-redundant idea units relative to the team's prior work                           | Surface difference mistaken for useful novelty               |
| Creativity        | Solution diversity (.30)    | Semantic and functional diversity among alternatives considered                      | Many weak alternatives mistaken for productive exploration   |
| Creativity        | Lexical originality (.25)   | Originality of expression after task-vocabulary adjustment                           | Language proficiency or style mistaken for creativity        |

## 4. Method

### 4.1 Expert Refinement Through Delphi

The expert panel included 12 members of whom four were from the educational assessment field, four from the field of educational data mining/learning analytics, and four from the field of interdisciplinary PBL, all of whom were purposely selected. Requirements included relevant research or assessment experience and at least 10 peer reviewed publications over the last 10 years. Round 1 included open questions to solicit opinions on candidate indicators, sources of evidence, validity threats, and governance needs. In rounds 2 and 3, participants were provided with anonymous peer feedback on their group's responses, and then reconsidered the importance of each item on the 0-1 scale. The interquartile range (IQR) of .50 or less was defined as consensus, and indicator weights were subsequently normalised within each competency so they sum to 1.00. This process is similar to the iterative feedback process used in Delphi research (Hasson et al., 2000). The study record was not archived with the recruitment method, the number of recruits per round, attrition or complete distributions by item. The information is therefore flagged as limitations of reporting and not as information that is assumed. A replication should keep a record of the

participants' flows and round summaries of the items in accordance with CREDES guidance (Jünger et al., 2017).

Consensus establishes agreement among this purposive expert panel; it does not demonstrate population representativeness, construct validity, or optimality of the weights. For that reason, the weights were treated as provisional parameters to be evaluated through external relations and sensitivity analysis.

## **4.2 Agent-Based Proof-of-Concept Simulation**

Because the retained record of the earlier proof of concept lacked executable code, seeds, row-level outputs, and complete metric definitions, a new simulation was specified independently rather than reconstructed. The structural design retained 500 synthetic learner agents, five assessment cycles, 50 replications, three competencies, and nine Delphi weights; previously reported simulation values were not used as calibration targets, stopping rules, or acceptance criteria. The model was documented with the Overview, Design concepts, and Details (ODD) protocol to support transparent replication (Grimm et al., 2020).

### ***4.2.1 Entities, Contexts, and Evidence Generation***

Each synthetic learner agent carried three time-varying latent competency states - critical thinking, collaboration, and creativity - across five assessment cycles. Latent competencies were generated identically for two synthetic observation-opportunity groups. Group B was assigned lower digital observability through a prespecified observation shift and additional missingness; the groups are modelling devices and do not represent real demographic categories. In each cycle, task complexity, collaboration support, and stress took low, medium, or high values, yielding 27 contextual combinations. Three observed indicators per competency were generated with bounded measurement noise and missingness, normalised to [0, 1], and aggregated with the Delphi weights. This structure separates the simulated competency state from the imperfect digital evidence available to the assessment system.

### ***4.2.2 Governance Conditions and Execution Sequence***

The system first calculated the weighted evidence coverage and a provisional composite score for each learner, competency, and cycle. A case was flagged if weighted evidence coverage was less than .75, the uncertainty proxy was greater than .11, there was no evidence available, or the batch coverage monitor identified a gap in the batch (group) greater than .03. The full-governance arm then provided structured human review for no more than 25% of cases per competency and cycle. A provisional estimate might be confirmed, amended, rejected or deferred, and flagged cases that were not resolved were explicitly not given a numeric score. For each numeric decision that was released there was an explanation record that had to be linked to it, and a disposition that had to be kept in an audit log. This primary comparison was made using matched synthetic learners and held constant latent trajectories, contexts, indicator draws, reviewer draws, and the provisional

scoring rule while the governance functions varied. The no-governance arm was programmed to accept all the computable provisional scores and to not include structured review, the batch fairness monitor, governance-triggered abstention, or explanation. The fairness monitor, explanations, and/or structured review were removed, one at a time, in three additional ablations. This matched arm model isolates the contribution and conflicting effects of governance components in the simulation from what would happen in real students, and it does not estimate causal effects in real students.

#### ***4.2.3 Replication; Metrics;***

Technical Reproducibility Production run consisted of 50 cycles with 500 agents each cycle. The model specification, parameters, outcomes, ablations, sensitivity ranges and date-coded seed schedule were set in stone prior to execution. Predictive agreement was the correlation of final and latent synthetic competency; mean absolute error (MAE) was a measure of score error; temporal fidelity was the comparison of observed and latent changes between adjacent cycles; and decision and explanation coverage was a measure of operational output. Fairness composite was calculated as 1 minus the average absolute gaps in MAE, false negative rates at a .65 proficiency level and decision coverage. The composite retained component gaps due to the ability of the composite to mask opposing effects. The accompanying package consists of ODD specifications, executable code, parameter file, exact seeds, row level synthetic data, summaries, sensitivity analyses, software version, and file hashes (Grimm et al., 2020).

#### **4.3 Pilot Field Study Participants, setting, and ethics.**

**4.3.1 Setting, participants and ethics**

The pilot took place in Semester 2, 2023-2024, at one research active university that already has a PBL provision across subjects. Forty-seven undergraduate students were involved in three courses and five faculty members supervised the projects. Twenty-four students got assessment reports in the framework, while 23 students continued with the existing process using the rubric. The recorded allocation system and course-level descriptors were not preserved to model clustering, discipline, team size or differential opportunities, and insufficient course-level descriptors were recorded to provide adequate digital traces. Thus, only a descriptive report of membership in a group is reported here, and the present analysis is not intended to make any causal or comparative effectiveness claims. The field study was approved by the institutional ethics committee, the protocol number of which has been removed for double-blinding. Written informed consent was obtained and participants were informed that there was no requirement that participation would

impact academic standing. The data was not accessible to other people except the authorised research team.

#### **4.3.2 Procedure and Measures**

Competency evidence was collected across a 10-week project cycle. The field instantiation generated Delphi-weighted composite scores from the nine normalised indicators; it did not fit a Random Forest or Gradient Boosting model to the 47-person sample. Framework reports displayed provisional competency scores and feature-based explanations. At project completion, framework scores were compared with holistic instructor ratings that, according to the archived record, were completed without access to the framework scores. Instructor ratings used project rubrics informed by established principles for analytic criteria and performance-level descriptions (Jonsson & Svingby, 2007). A second assessor rated a random 30% of project outputs. Because the archived record does not specify the ICC model, unit, or confidence interval, the reported ICCs are treated as descriptive source values rather than independently reproducible reliability estimates.

#### **4.4 Statistical and Evidentiary Strategy**

Pearson correlations quantified relations between framework scores and instructor ratings. Two-sided 95% confidence intervals were reconstructed with Fisher's z transformation using  $n = 47$ . Correlation supports a preliminary convergent-validity argument, but it does not establish score interchangeability, predictive validity, fairness, or absence of shared-method influences (Campbell & Fiske, 1959). Omega coefficients and ICCs were retained at their reported values because row-level pilot data were unavailable for independent recalculation. No subgroup fairness analysis was conducted because the sample did not provide adequate cell sizes.

The earlier study record included an inconsistent power statement for an independent-groups effect. For a two-sided test with  $\alpha = .05$ , 80% power, and  $d = .50$ , approximately 128 participants are required in total. With  $n_1 = 24$  and  $n_2 = 23$ , power to detect  $d = .50$  is approximately .39, and an effect of about  $d = .84$  is required for 80% power (Faul et al., 2007). Group-effect and moderation claims were therefore removed. This correction does not alter the Delphi or pilot correlation results; it narrows the claims supported by the field pilot. Simulation results are reported separately as outputs of the new synthetic model.

**Table 4**

#### *Evaluation Components and Permissible Inferences*

| Component | Unit and analysis                                                    | Evidence provided                         | Inference not supported                           |
|-----------|----------------------------------------------------------------------|-------------------------------------------|---------------------------------------------------|
| Delphi    | 12 experts, three rounds;<br>IQR consensus and<br>normalised weights | Content relevance and expert<br>agreement | Optimal or universally valid<br>indicator weights |

| Component         | Unit and analysis                                                                                                       | Evidence provided                                                                                           | Inference not supported                                                                              |
|-------------------|-------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| Simulation        | 500 agents, five cycles, 50 replications; frozen new specification, matched arms, 27 contexts, and sensitivity analysis | Reproducible operational behaviour, component trade-offs, and governance ablations under stated assumptions | Real-world effectiveness, protected-group fairness, or causal superiority beyond the synthetic model |
| Pilot convergence | 47 students; Pearson r and Fisher 95% CI                                                                                | Preliminary relation to independent instructor judgement                                                    | Interchangeability, causal effects, or generalisability                                              |
| Pilot omega       | Three indicators per composite; source-reported omega                                                                   | Descriptive coherence under the reported scoring model                                                      | Full reliability evidence for formative composites                                                   |
| Pilot ICC         | Random 30% double-rated; model unspecified                                                                              | Encouraging source-reported agreement                                                                       | Replicable inter-rater reliability without model and interval                                        |

## 5. Results

### 5.1 Delphi Refinement

By Round 3, all nine indicators met the prespecified IQR threshold. For critical thinking, Inquiry depth and Source diversity each received a weight of .35 and Substantive revision .30; the reported IQRs were .20, .30, and .20, respectively. For collaboration, Task balance received .40, Interaction frequency .35, and Response timeliness .25; all reported IQRs were .30 or less. For creativity, Novelty rate received .45, Solution diversity .30, and Lexical originality .25; all reported IQRs were likewise .30 or less. The panel therefore placed the largest relative weight on balanced participation for collaboration and on ideational novelty for creativity.

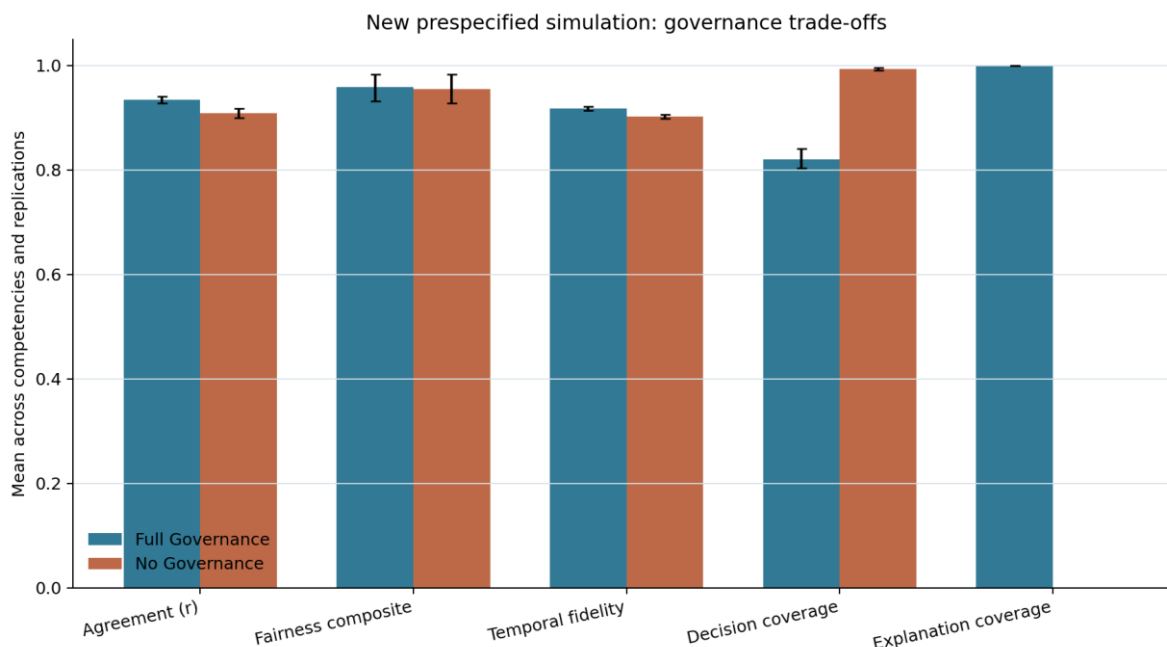
### 5.2 Simulation Proof of Concept

Across 50 replications, all 500 agents completed five synthetic cycles in every model arm. Under full governance, predictive agreement with latent competency was  $r = .934$  for collaboration and  $r = .935$  for both creativity and critical thinking; MAE was .059 for each competency, temporal fidelity ranged from .918 to .919, and numeric decision coverage was .820. All numeric full-governance decisions carried an explanation record. The matched no-governance arm achieved agreement of  $r = .908$ -.909 and MAE of .070, with decision coverage of .993 but no explanation records. Paired replication differences favoured full governance for agreement (.025-.026) and MAE (-.011 to -.012) but showed a .173 reduction in decision coverage because unresolved flagged cases abstained. The mean fairness composite was .955-.962 under full governance and .953-.958 without governance; paired Monte Carlo intervals crossed zero, so the simulation does not establish a stable overall fairness advantage.

**Table 5***New Agent-Based Simulation Results by Competency and Governance Arm*

| Competency        | Arm             | Agreement r       | MAE               | Fairness composite | Temporal fidelity | Decision coverage |
|-------------------|-----------------|-------------------|-------------------|--------------------|-------------------|-------------------|
| Critical thinking | Full governance | .935 [.928, .941] | .059 [.057, .061] | .955 [.927, .980]  | .919 [.914, .922] | .820 [.806, .845] |
| Critical thinking | No governance   | .909 [.899, .919] | .070 [.068, .072] | .952 [.929, .977]  | .903 [.899, .906] | .993 [.990, .996] |
| Collaboration     | Full governance | .934 [.930, .939] | .059 [.058, .061] | .962 [.929, .985]  | .918 [.914, .922] | .820 [.804, .841] |
| Collaboration     | No governance   | .909 [.901, .918] | .070 [.068, .073] | .958 [.931, .988]  | .903 [.898, .907] | .993 [.991, .996] |
| Creativity        | Full governance | .935 [.927, .941] | .059 [.057, .061] | .961 [.938, .986]  | .919 [.915, .923] | .820 [.801, .835] |
| Creativity        | No governance   | .908 [.900, .917] | .070 [.068, .073] | .955 [.925, .985]  | .903 [.899, .906] | .993 [.990, .996] |

**Note.** Values are means with empirical 2.5th-97.5th percentile Monte Carlo intervals across 50 replications. These intervals describe variation under the synthetic model, not uncertainty in a real student population. MAE = mean absolute error. Component fairness gaps and sensitivity analyses are reported in the supplementary package.

**Figure 2***Primary Governance Trade-offs in the New Prespecified Simulation*

**Note.** Bars show macro-averaged means across the three competencies; error bars summarise empirical 2.5th-97.5th percentile ranges across 50 replications. Explanation coverage denotes the presence of a linked record, not explanation quality. The lower decision coverage under full governance reflects explicit abstention when required review exceeded the fixed capacity.

### 5.3 Pilot Measurement Evidence

Framework scores were positively related to independent holistic instructor ratings. The reconstructed results were  $r = .71$ , 95% CI [.53, .83], for critical thinking;  $r = .74$ , 95% CI [.58, .85], for collaboration; and  $r = .79$ , 95% CI [.65, .88], for creativity. The intervals exclude small associations and are consistent with substantial convergence, while still leaving uncertainty about the exact population relationships.

The explanation ablation changed explanation coverage from 1.00 to 0 while leaving scores unchanged, as expected from its documentary role. Removing structured review reduced decision coverage to approximately .570 and widened the group decision-coverage gap to approximately .204-.214. Removing the batch fairness monitor produced no numerical change under the frozen thresholds because the individual evidence and uncertainty rules had already flagged the same incomplete cases. This null ablation result is retained rather than concealed or followed by post hoc threshold tuning. Table 5 presents the primary simulation results, and Figure 2 shows the principal governance trade-offs. Reported pilot omega coefficients were .78 for critical thinking, .81 for collaboration, and .76 for creativity; reported ICCs were .83, .79, and .76, respectively.

These pilot coefficients remain provisional and source-reported. Table 6 consolidates the pilot evidence without treating the coefficients as equivalent indicators of validity.

**Table 6**

*Pilot Measurement Results (N = 47)*

| Competency        | Correlation with instructor rating | 95% CI     | Omega | Reported ICC | Interpretation                                             |
|-------------------|------------------------------------|------------|-------|--------------|------------------------------------------------------------|
| Critical thinking | .71                                | [.53, .83] | .78   | .83          | Substantial preliminary convergence; descriptive coherence |
| Collaboration     | .74                                | [.58, .85] | .81   | .79          | Substantial preliminary convergence; descriptive coherence |
| Creativity        | .79                                | [.65, .88] | .76   | .76          | Strongest observed convergence; descriptive coherence      |

## 6. Discussion

The study devised a governance-integrated AI architecture, and engaged in evaluation dynamically appropriate to various components of this artefact. Nine provisional indicators of the content and relative weighting were agreed by experts. The workflow was found to be repeatable and could be used to generate scores, run governance checks, and provide explanations to repeated synthetic cycles during the simulation. The weighted composites were related to independent instructor judgements as initial evidence provided by the field pilot. These findings together indicate the need for further development and a larger study to validate. They are not validations, validations of the intervention effectiveness, or fairness of intervention in real classrooms of students.

### 6.1 Theoretical contribution – From model accuracy to a coupled validity system

The key contribution is the adoption of a change in the unit of analysis. An AI-supported assessment is not just a 'map' from features to scores; it's a sociotechnical measurement system where evidence takes on meaning, score is created, consequences are taken into account, authority is exercised and errors are corrected. This perspective merges the concepts of educational measurement and human-centred AI. There are constraints on data that can be gathered and what can be concluded from the data; the analysis of fairness relies on the construct and on the planned use of the data; explanation is only meaningful if there is the opportunity to review this explanation; human governance is meaningful only if the person reviewing it has evidence, time, and authority.

The way these functions relate to each other is described by a controlled coupling. A standard pipeline fetches the assumptions at the moment the model is built and passes the results forward. The proposed architecture allows for invalidate assumption with later information. A reasoned override by an educator can identify a missing context variable, repeated appeals can identify a biased feature, and a rubric revision can result in a semantic version change on the target construct. Stopping and safely decoupling is an important skill as well. A protected group check is less defensible than forced scoring when evidence is not available, or when a protected group check is unstable, or when an explanation is not consistent with project evidence.

The framework also defines governance integration and governance centralisation. Common definitions, traceable handoffs, reciprocal updating, and documented responsibility are all key components of integration. It doesn't need a single platform or open data transfers. A function can be modular and access-controlled, as long as the versioned interfaces maintain the connections between evidence, scoring, review and action.

### 6.2 Interpretation of the Pilot Evidence

The correlations show that there was a fair amount of common variance between the digital composites and the instructor ratings. The non-unit correlation is anticipated since the methods rely on different strands of evidence: composites on documented longitudinal evidence, and instructors on contextual or pedagogical evidence. That the two numbers are different should not be called “desirable.” It simply is an observation that needs to be explained. Future research to test multi-trait-multimethod studies is needed to determine if each competency is most highly associated with its respective external criterion, and to explore cases of disagreement qualitatively.

With regard to the omega results, they must be taken with great care. High internal consistency is generally considered to be reliability, but this assumption is that the items are the effects of a common latent variable (a reflective model). Here a variety of components were intentionally integrated to illustrate a wide range of performance. A more robust validation programme would consider the stability of the programme over programming periods, sensitivity to different weights, generalizability across tasks and raters, relationships with external performance measures, and consistency of decisions across realistic data perturbations. Omega could still be used as a supplementary argument of reliability but it should not be the primary argument (Bollen & Lennox, 1991; McNeish, 2018).

### 6.3 Governance and Fairness Implications

Use of the fairness composite in the simulation should not be confused with the fact that the framework is fair in practice. It was intentionally made up of two groups, latent competency was created in both groups, and the opportunity disparity was assumed. The reduction in MAE and false negatives was offset by the growth of decision coverage gaps, as the decision coverage gap widened due to capacity constrained abstention, with a relatively small and inconsistent advantage found across replications when considering the composite advantage. This trade-off illustrates the need for a complement, such as a component measure and coverage, to any fairness composite. The institutions have to define protected and educationally significant groups, explain error criteria and thresholds, publish the results disaggregated with marginal error and report with uncertainty, and avoid penalising students who are doing their work offline.

Human review is not presumed to be effective, it must be assessed. Review time, agreement and disagreement pattern, reasons for disagreement or changes, correction rate, student understanding, appeal result and identification of seeded errors should be measured in further research. Intended use, excluded uses, training and evaluation data, subgroup results, known limitations, and changes across versions may be recorded in model cards and end-to-end audit records (Mitchell et al., 2019; Raji et al., 2020).

### 6.4 Practical Recommendations for Institutional Adoption

**Recommendation 1 - Complete a construct-and-data audit before selecting a model.** For every indicator, institutions should document the competency claim, evidence source, extraction and normalisation rule, missing-data treatment, plausible proxy threat, differential opportunity to

produce the evidence, retention period, and a less intrusive alternative. Indicators that cannot be connected to a defensible score interpretation should not enter the system (American Educational Research Association et al., 2014; Jacobs & Wallach, 2021; Tabassi, 2023).

**Recommendation 2 - Begin in shadow mode and restrict use to formative or low-stakes decisions.** Reports should initially be generated for evaluation without determining grades, progression, or access to opportunities. Institutions should compare system outputs with independent educator judgement, investigate disagreements, and prespecify the evidence required before any expansion of use. The present study does not support fully automated high-stakes assessment (European Parliament & Council of the European Union, 2024; UNESCO, 2023).

**Recommendation 3 - Resource human governance as an operational function.** Institutions should identify who may view evidence, confirm or override a score, suspend processing, and decide an appeal. Reviewer capacity must be tested against expected flag rates; when capacity is exceeded or evidence is insufficient, the system should abstain and record the reason rather than issue a forced score. Review time, override reasons, correction rates, unresolved-case rates, and appeal outcomes should be monitored (Green, 2022; Raji et al., 2020).

**Recommendation 4 - Monitor fairness with disaggregated component measures.** Before consequential use, institutions should define educationally relevant and legally protected groups, justify comparison criteria and thresholds, report MAE, false-negative, and decision-coverage gaps with uncertainty, and examine whether missing digital traces reflect offline work or unequal access. A single fairness composite should never substitute for the component results or their educational interpretation (Baker & Hawn, 2022; Mitchell et al., 2021).

**Recommendation 5 - Provide usable explanations, notice, and contestation.** Educators need training in construct interpretation, uncertainty, feature attribution, automation bias, and documented override. Students should receive plain-language notice of the data used, the intended and excluded uses, the limits of the score, how to inspect or correct evidence, how to request reconsideration, and who remains accountable for the final judgement (Khosravi et al., 2022; Mitchell et al., 2019; UNESCO, 2023).

**Recommendation 6 - Authorise expansion only through staged evidence gates.** Movement beyond shadow or low-stakes use should require multi-institutional validation, learner-level data partitioning, adequate protected-group precision, privacy and security review, accessible appeal, versioned documentation, and an independent approval decision for the specific intended use. Material changes to constructs, features, thresholds, or models should trigger renewed validation rather than silent continuation (Raji et al., 2020; Tabassi, 2023).

## 6.5 Limitations and Research Agenda

There are a number of limitations to the study. First, the field pilot only included 47 students at one institution and did not consider clustering by course, instructor, or team. Second, the

replacement simulation is well-defined and can be replicated, but it is a simulation and thus abstract; the results depend upon specific assumptions about latent dynamics, observation shifts, missingness, reviewer noise, thresholds, and capacity. Sensitivity analysis can be used to determine parameter dependence but cannot provide nuggets of empirical realism. Third, the fairness analysis does not involve any protected groups, and the null fairness-monitor ablation result demonstrates that rules may be sufficient to avoid the necessity of fairness monitoring under certain circumstances. Fourthly, the framework is based on digital representations that may not capture the face-to-face discussion, informal mentoring, embodied design work, or unequal contexts of access. Fifth, protected-group fairness would not be assessed effectively during the pilot. Sixth, the formative nature of the composites restricts interpretation of omega. Seventh, the ICC model and interval estimates were not available and there was no independent system for retaining row-level pilot and Delphi data for reconstruction. These constraints lead to a stage-based program. In stage 1, the measurement specification should be finalized, which includes creating maps, deriving rules for extracting features, filling in missing data, interrater protocols, sensitivity to weights, and cognitive interviews about the explanation. In stage 2, a multi-institutional shadow deployment with learner-level data partitions and course/team random effects should be conducted. To monitor both benefits and harms, stage 3 should evaluate the same scoring model in the presence of government structure ablations (prespecified) including no explanation, no fairness review, or no structured override. Stage 4 should consider the consequences such as workload, appeal, trust, subgroup outcomes and changes in teaching practice. Confirmatory claims should be submitted with Code, synthetic data, definitions of metrics, and analytic scripts. 7.

## 7. Conclusion

This study reports an architecture to evaluate interdisciplinary PBL's critical thinking, collaboration, and creativity, coupled with a governance framework. The main idea is that the process of responsible AI assessment is a "coupled validity" process: longitudinal evidence, provisional scores, fairness and explanation checks, and authorised human judgement should all have a bidirectional flow of information and be open to review. Provisional indicators and weights were set with expert consensus, a newly specified, reproducible simulation showed better synthesized agreement, as well as reduced error and a significant abstention trade-off, and a small field pilot produced significant correlations with independent instructor assessments. The evidence is promising but preliminary. It supports transparent replication, construct-and-data audit, shadow deployment, documented human review, disaggregated fairness monitoring, accessible appeal, and staged authorisation. It does not support real-world fairness claims or automated high-stakes use.

## Data Availability Statement

The complete replacement-simulation package - frozen analysis plan, ODD model specification, executable code, parameter file, exact seeds, feature specification, data dictionary, compressed

row-level synthetic data, replication and scenario summaries, sensitivity analyses, figures, software versions, and SHA-256 file hashes - accompanies this manuscript. Before submission, the placeholder below must be replaced with the active anonymised OSF view-only URL: [ANONYMISED OSF VIEW-ONLY LINK]. Pilot participant data remain subject to institutional ethics and privacy restrictions; de-identified summary data and a data dictionary may be shared only under an approved data-use agreement.

## Declarations

**Ethical approval and consent to participate.** The pilot field study received institutional ethics approval (protocol reference redacted for double-blind review). All participants provided written informed consent.

**Competing interests.** The authors declare no competing interests.

**Funding.** Funding information is redacted for double-blind review.

**Use of AI-assisted tools.** An AI-assisted coding and language environment was used to help draft and debug simulation code, prepare technical documentation, edit language, and format references. The simulation specification was frozen before the production run, legacy simulation values were not used as calibration targets, and the authors reviewed the model assumptions, executed and validated the code, checked the outputs, verified the sources, interpreted the findings, and approved the final manuscript. The AI system is not an author and bears no responsibility for the work.

## References

- Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association. <https://www.testingstandards.net/open-access-files.html>
- 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/s40593-021-00285-9>
- Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110(2), 305-314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(Suppl. 3), 7280-7287. <https://doi.org/10.1073/pnas.082080899>

- Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, Article 4.  
<https://doi.org/10.1186/s41239-023-00436-z>
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.  
<https://doi.org/10.1037/h0046016>
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163. <https://doi.org/10.1089/big.2016.0047>
- 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). Association for Computing Machinery.  
<https://doi.org/10.1145/2090236.2090255>
- European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191. <https://doi.org/10.3758/BF03193146>
- Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64-71. <https://doi.org/10.1007/s11528-014-0822-x>
- Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 105681.  
<https://doi.org/10.1016/j.clsr.2022.105681>
- Grimm, V., Railsback, S. F., Vincenot, C. E., Berger, U., Gallagher, C., DeAngelis, D. L., Edmonds, B., Ge, J., Giske, J., Groeneveld, J., Johnston, A. S. A., Milles, A., Nabe-Nielsen, J., Polhill, J. G., Radchuk, V., Rohwäder, M.-S., Stillman, R. A., Thiele, J. C., & Ayllón, D. (2020). The ODD protocol for describing agent-based and other simulation models: A second update to improve clarity, replication, and structural realism. *Journal of Artificial Societies and Social Simulation*, 23(2), Article 7.  
<https://doi.org/10.18564/jasss.4259>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315-3323.  
<https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>

- Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, 32(4), 1008-1015.  
<https://doi.org/10.1046/j.1365-2648.2000.t01-1-01567.x>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75-105. <https://doi.org/10.2307/25148625>
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32(3), 504-526.  
<https://doi.org/10.1007/s40593-021-00239-1>
- Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: A systematic review. *Educational Technology Research and Development*, 68(4), 1961-1990. <https://doi.org/10.1007/s11423-020-09788-z>
- Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375-385). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445901>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144.  
<https://doi.org/10.1016/j.edurev.2007.05.002>
- Jünger, S., Payne, S. A., Brine, J., Radbruch, L., & Brearley, S. G. (2017). Guidance on Conducting and REporting DELphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. *Palliative Medicine*, 31(8), 684-706. <https://doi.org/10.1177/0269216317690685>
- Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1-33.  
<https://doi.org/10.1007/s10115-011-0463-8>
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (Article 43)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.  
<https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
- Kokotsaki, D., Menzies, V., & Wiggins, A. (2016). Project-based learning: A review of the literature. *Improving Schools*, 19(3), 267-277.  
<https://doi.org/10.1177/1365480216659733>
- Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, 3, 100074.  
<https://doi.org/10.1016/j.caeai.2022.100074>

- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774. <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412-433. <https://doi.org/10.1037/met0000144>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>
- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (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>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45-77. <https://doi.org/10.2753/MIS0742-1222240302>
- Perignat, E., & Katz-Buonincontro, J. (2019). STEAM in practice and research: An integrative literature review. *Thinking Skills and Creativity*, 31, 31-43. <https://doi.org/10.1016/j.tsc.2018.10.002>
- Rababah, L. M. (2022). Contextualization to enhance students' writing ability. *Theory and Practice in Language Studies*, 12(11), 2316-2321.
- Rababah, L. M. (2023). Stepping into language mastery: Virtual reality simulations as catalysts for EFL pronunciation enhancement. *Journal of Applied Learning and Teaching*, 6(Special Issue 1), 110-122.
- Rababah, L. M., Al-Shboul, O. K., & Ahmed Banikalef, A. A. (2023). Examination of the use of feedback in EFL writing instruction: A case study of Jordan. *Journal of Language Teaching and Research*, 14(1), 263-268.
- Rababah, L. M., & Rababah, M. A. I. (2024). Online social learning and instructional presence: Enhancing English education at a Jordanian university. *Forum for Linguistic Studies*, 6(6), 729-741.
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372873>
- 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). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939778>
- Suresh, H., & Gutttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Article 17, pp. 1-9). Association for Computing Machinery. <https://doi.org/10.1145/3465416.3483305>
- Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- UNESCO. (2023). Guidance for generative AI in education and research. <https://doi.org/10.54675/EWZM9535>
- Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98-110. <https://doi.org/10.1016/j.chb.2018.07.027>
- Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences. *Journal of Curriculum Studies*, 44(3), 299-321. <https://doi.org/10.1080/00220272.2012.668938>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education - Where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>