



## Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence

**Basile Mulwani Makelele**

Full Professor at University of Lubumbashi (UNILU), Democratic Republic of the Congo

**ABSTRACT:** The rapid expansion of generative artificial intelligence (GenAI) is changing the conditions under which teaching, learning evidence, and scientific research are produced and assessed in higher education. This study examines these transformations through a qualitative-quantitative documentary analysis of 686 publications published between 2022 and 2026. Each document was coded using 44 binary thematic indicators grouped into six domains: assessment of teaching, assessment of learning, research supervision, assessment by academic juries, GenAI-related risks and limitations, and assessment transformations or good practices. Descriptive thematic frequencies were complemented by Multiple Correspondence Analysis (MCA), Benzécri and Greenacre corrections of inertia, and pairwise Phi coefficients. Publications from 2025 and 2026 accounted for 72.0% of the corpus, confirming the very recent and rapidly expanding character of the field. After Benzécri correction, the first three MCA dimensions explained 70.20% of corrected inertia, compared with 21.74% of raw inertia; the more conservative Greenacre correction yielded 47.86%. Phi analysis showed a strong association between authentic assessment and oral/defence-based assessment ( $\phi = .694$ ;  $\chi^2 = 330.43$ ;  $p < .001$ ), while process traceability was positively associated with human validation ( $\phi = .177$ ;  $\chi^2 = 21.41$ ;  $p < .001$ ). Across the corpus, integrity, transparency, explicit task-level policies, AI literacy, authentic assessment, traceability, oral justification, and human validation emerged as recurrent responses. The findings indicate a shift from assessment centred on the final product toward an evidence-based ecosystem combining product, process, traceability, explicit justification, dialogue, and accountable human judgment. On this basis, the study proposes the P-P-T-E-V model—Product, Process, Traceability, Explication, and Human Validation—as an integrative framework for assessing teaching-related evidence and scientific research in the GenAI era.

**KEY WORDS:** generative artificial intelligence, higher education assessment, scientific research, research supervision, multiple correspondence analysis, academic integrity

**Cite the Article:** Makelele, B.M. (2026). *Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence*. *Contemporary Research Analysis Journal*, 3(10), 659-670. <https://doi.org/10.55677/CRAJ/06-2026-Vol03I10>

*License:* This is an open-access article under the CC BY 4.0 license: <https://creativecommons.org/licenses/by/4.0/>

*Publication Date:* October 08, 2026

*\*Corresponding Author:* Mulwani Makelele Basile

### I. INTRODUCTION

Generative artificial intelligence (GenAI) is among the fastest technological transformations to affect higher education. Tools capable of generating text, synthesizing documents, answering complex questions, commenting on academic work, producing computer code, and assisting scientific analysis are altering the relationship between learning, intellectual production, and assessment.<sup>14,30</sup> This development is no longer simply about access to a new digital tool. It directly affects the conditions under which universities can establish that a student, researcher, or candidate has actually mastered the knowledge, skills, reasoning, and scientific procedures attributed to them.

The central problem therefore extends beyond whether students or researchers use AI. It concerns the validity of the assessment inference: to what extent does a product submitted to a lecturer, supervisor, or academic jury still constitute sufficient evidence of its author's competence? This question becomes particularly important when generative systems can produce outputs that display the formal characteristics traditionally associated with high-quality academic work.<sup>22,13</sup>

Real-world evidence illustrates the problem. In a university examination system, fully AI-generated answers were inserted into authentic assessment conditions; most escaped detection and achieved high marks, challenging the assumption that the final written product alone reliably identifies individual competence.<sup>22</sup> Comparable concerns arise from studies showing that ChatGPT can perform competitively on discipline-specific assessments in physiology and accounting.

## Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence

In accounting education, a large international comparison showed that ChatGPT could answer a substantial proportion of assessment questions, although students still outperformed it overall.<sup>29</sup> In physiology, ChatGPT outperformed a sample of medical students on a university examination.

These findings have encouraged a shift away from responses based exclusively on surveillance or detection toward redesigning assessment itself.<sup>13,30</sup> Proposed strategies include diversifying evidence, using contextualized tasks, documenting the production process, adding oral or dialogic components, and explicitly defining which forms of AI assistance are permitted for each task.

This transformation affects at least three interdependent university functions: assessment of teaching, assessment of learning, and assessment of scientific research. In the research domain, supervision must also be distinguished from jury assessment. Supervision supports the progressive construction of scientific competence, whereas the jury must determine whether the candidate actually understands the research problem, methodological choices, analyses, interpretations, limitations, and contribution.

Automation does not reduce the issue to a simple opposition between human and algorithmic grading. Research on large language models in feedback and grading shows useful performance in some contexts, but also sensitivity to prompts, criteria, disciplinary context, and atypical responses. Dai et al. evaluated the capacity of GPT models to generate assessment feedback, while Flodén compared ChatGPT grading with human grading of higher-education examinations.<sup>4,7</sup> Other work indicates that students' perceptions of feedback can change when they know that it was produced by AI.

The same pattern is visible in formative feedback. Large language models can produce rapid and relevant comments when their use is guided by teachers, yet they may struggle with atypical cases and with highly response-specific feedback.<sup>5,6</sup> The emerging position is therefore one of complementarity rather than systematic replacement of human academic judgment.

### From the final product to a bundle of evidence

University assessment traditionally uses an observable product as evidence of a competence that cannot be observed directly. An essay is used to infer argumentative ability; an examination to infer knowledge; a thesis to infer research competence; and a defence to infer the ability to understand, explain, and justify a scientific process. GenAI weakens the link between the observable product and the competence attributed to its author when the final output may result from substantial human-AI collaboration.

A coherent text no longer demonstrates, by itself, that the student constructed the reasoning it contains. A formally correct bibliography does not guarantee that the references were verified. A polished analysis does not always reveal how the data were processed. Assessment therefore requires several converging indicators: the final product, evidence from the process, justification of decisions, interaction with the evaluator, and, when appropriate, oral defence.

Institutional responses are evolving in this direction. Analyses of university guidance have emphasized academic integrity, assessment redesign, communication with students, and increasingly regulated acceptance rather than blanket prohibition of GenAI.<sup>14,12</sup> The Artificial Intelligence Assessment Scale (AIAS) similarly shifts attention from the binary question of whether AI is allowed toward the level and type of assistance permitted for a specific task.

The AIAS and its early implementation studies provide a graded framework for aligning permitted AI use with intended learning outcomes.<sup>18,9</sup> This distinction is especially important for scientific research, where a thesis or dissertation results from a chain of intellectual operations: problem formulation, literature search, methodological decisions, data collection or selection, analysis, interpretation, and scientific argumentation.

Accordingly, assessment of scientific research in the GenAI era should not be limited to the formal quality of the final manuscript. It should also consider the traceability of the intellectual process, the researcher's capacity to justify decisions, verify sources, reproduce or explain analyses, and defend the logic of the work. The fundamental question is therefore not simply whether AI was used, but what the student or researcher can actually understand, perform, verify, explain, and defend.

### Study objective

The general objective of this study is to analyse transformations in the assessment of teaching and scientific research in higher education in the era of GenAI, while treating assessment of learning as a core source of evidence about teaching quality and student competence. Specifically, the study aims to identify dominant themes in the literature; characterize transformations affecting teachers, students, research supervisors, and academic juries; identify multidimensional thematic structures through MCA; examine associations between indicators using Phi coefficients; and propose an integrative assessment framework for the GenAI era.

## II. METHOD

### Study design and documentary corpus

The study used a documentary, descriptive, analytical, and exploratory design combining qualitative content analysis with quantitative multivariate analysis. This combination is appropriate for a rapidly evolving field in which systematic reviews, bibliometric analyses, content analyses, and empirical studies coexist.<sup>2,8,19,21</sup>

The documentary database was structured by record number, thematic axis, authors, year, topic, objectives, research questions or hypotheses, methodology, sample or corpus, and main findings. The corpus comprised 686 documents published between 2022 and

## Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence

2026: 1 document in 2022, 56 in 2023, 135 in 2024, 242 in 2025, and 252 in 2026. Thus, 494 publications (72.0%) were published in 2025–2026, a distribution consistent with the rapid expansion of GenAI scholarship in higher education.<sup>8,21</sup>

### Unit of analysis and thematic coding

Each document constituted one unit of analysis. Available fields were subjected to systematic lexical-thematic coding. Six thematic domains were retained: assessment of teaching; assessment of learning; scientific research from the supervisor's perspective; scientific research from the jury's perspective; GenAI-related risks and limitations; and assessment transformations or good practices. Forty-four binary indicators were constructed, with presence coded 1 and absence coded 0.

The quantitative coding used in this study was systematic lexical-thematic coding; it should not be interpreted as independent double manual coding. This distinction is methodologically important. Studies using multiple coders should document the coding procedure and, where appropriate, inter-rater agreement. The absence of independent double coding is therefore treated as a limitation rather than concealed as a reliability procedure that was not performed.

### Descriptive analysis

Frequencies were calculated for each thematic category. Because the thematic tables are multiple-response tables, one publication could be coded under several indicators; percentages within a table are therefore not expected to sum to 100%. The denominator for each indicator remained  $N = 686$ . This descriptive stage identified dominant themes before the analysis of their multivariate organization.

### Multiple Correspondence Analysis

Multiple Correspondence Analysis (MCA) was applied to the  $686 \times 44$  binary matrix to identify proximities and oppositions among modalities and to reveal the principal structures of the thematic space. The first five dimensions explained 8.05%, 7.63%, 6.07%, 5.65%, and 4.78% of raw inertia, respectively. Because a large number of dichotomous variables disperses inertia across many dimensions, interpretation did not rely on raw inertia alone. Benzécri and Greenacre corrections were therefore used to provide more informative estimates of dimensional importance.

With  $Q = 44$  variables, the Benzécri threshold was  $1/44 = 0.0227$ . The first three dimensions represented 30.64%, 26.33%, and 13.23% of Benzécri-corrected inertia, for a cumulative 70.20%. Under the more conservative Greenacre correction, the first three dimensions represented 20.89%, 17.95%, and 9.02%, for a cumulative 47.86%. MCA was used for exploratory structural interpretation rather than causal inference.

### Associations between thematic indicators

Associations among the 44 dichotomous indicators were examined using Phi coefficients ( $\phi$ ) with Pearson  $\chi^2$  tests at one degree of freedom. In total, 946 pairs ( $44 \times 43 / 2$ ) were examined. Four pairs represented categories coded identically across tables and mechanically produced  $\phi = 1$ ; these were excluded from substantive interpretation. Because of the large number of tests, p-values are treated cautiously. If the analysis is used confirmatorily, control of the false discovery rate, for example by the Benjamini-Hochberg procedure, is recommended.<sup>20</sup>

The analytical strategy therefore combined three complementary levels: thematic frequencies, exploratory multivariate structure through MCA, and bivariate association through Phi. This triangulation moves the analysis from isolated frequencies to the organization and co-occurrence of themes across the 686-document corpus.

## III. RESULTS

Results are presented at four complementary levels: (1) temporal and methodological characteristics of the corpus; (2) thematic frequencies relating to teaching, learning evidence, and scientific research; (3) the multidimensional structure identified by MCA; and (4) associations between binary indicators using Phi coefficients. For all multiple-response thematic tables,  $N = 686$ .

**Table 1: Distribution of documents by year of publication**

| Year  | f   | %     |
|-------|-----|-------|
| 2022  | 1   | 0.1   |
| 2023  | 56  | 8.2   |
| 2024  | 135 | 19.7  |
| 2025  | 242 | 35.3  |
| 2026  | 252 | 36.7  |
| Total | 686 | 100.0 |

Scientific production was strongly concentrated in the most recent years. Publications from 2025 (35.3%) and 2026 (36.7%) together accounted for 494 documents, or 72.0% of the corpus. The field is therefore both recent and rapidly expanding.

**Table 2: Distribution by methodological orientation**

| Methodological orientation | f   | %     |
|----------------------------|-----|-------|
| Other / not specified      | 221 | 32.2  |
| Quantitative               | 171 | 24.9  |
| Conceptual / critical      | 87  | 12.7  |
| Review / synthesis         | 79  | 11.5  |
| Qualitative                | 69  | 10.1  |
| Mixed methods              | 59  | 8.6   |
| Total                      | 686 | 100.0 |

Quantitative studies represented 24.9% of the corpus. Conceptual or critical analyses (12.7%), reviews or syntheses (11.5%), qualitative studies (10.1%), and mixed-methods studies (8.6%) indicate substantial methodological diversity. The large 'other/not specified' category (32.2%) requires caution when interpreting the methodological profile.

## Assessment of teaching

**Table 3: Content relating to the assessment of teaching**

| Identified content                               | f   | %    |
|--------------------------------------------------|-----|------|
| Pedagogical quality and effectiveness            | 2   | 0.3  |
| Design/scenario and pedagogical alignment        | 37  | 5.4  |
| Teacher practices and role                       | 133 | 19.4 |
| Feedback and pedagogical feedback                | 128 | 18.7 |
| Student engagement, satisfaction, and experience | 35  | 5.1  |
| Equity, inclusion, and accessibility             | 42  | 6.1  |

Assessment of teaching was most often represented through teacher practices and roles (19.4%) and feedback (18.7%). Pedagogical design and alignment (5.4%), equity and accessibility (6.1%), and student experience (5.1%) were less frequent. The explicit label 'pedagogical quality and effectiveness' was rare (0.3%), suggesting that teaching quality is more often operationalized through specific practices than named directly.

## Assessment of learning evidence

**Table 4: Content relating to the assessment of learning**

| Identified content                            | f   | %    |
|-----------------------------------------------|-----|------|
| Tasks, assignments, and written work          | 110 | 16.0 |
| Examinations, tests, and quizzes              | 152 | 22.2 |
| Performance and learning outcomes             | 75  | 10.9 |
| Authentic assessment and contextualized tasks | 59  | 8.6  |
| Rubrics, criteria, and grids                  | 28  | 4.1  |
| Automated grading/assessment                  | 26  | 3.8  |
| Process evidence and traceability             | 35  | 5.1  |

Examinations, tests, and quizzes were the most frequent learning-assessment category (22.2%), followed by written assignments (16.0%) and performance or learning outcomes (10.9%). Authentic and contextualized assessment appeared in 8.6% of documents. Process evidence and traceability accounted for 5.1%, while rubrics and automated grading were less frequent. Traditional formats remain central, but the literature increasingly considers mechanisms for authenticating competence.

## Scientific research: the supervisor's function

**Table 5: Assessment of scientific research from the supervisor's perspective**

| Identified content                        | f  | %   |
|-------------------------------------------|----|-----|
| Supervisory feedback and supervision      | 26 | 3.8 |
| Scientific writing and revision           | 49 | 7.1 |
| Literature search, sources, and citations | 35 | 5.1 |

## Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence

| Identified content                        | f   | %    |
|-------------------------------------------|-----|------|
| Methodology, analysis, and interpretation | 10  | 1.5  |
| Integrity, authorship, and attribution    | 162 | 23.6 |
| Human verification and validation         | 31  | 4.5  |

From the perspective of research supervision, integrity, authorship, and attribution clearly dominated (23.6%). Scientific writing and revision represented 7.1%, literature search and citations 5.1%, and human validation 4.5%. The corpus therefore places particular emphasis on intellectual responsibility in a context in which GenAI can intervene at several stages of research production.

### Scientific research: the jury's function

**Table 6: Assessment of scientific research from the academic jury's perspective**

| Identified content                                | f   | %    |
|---------------------------------------------------|-----|------|
| Scientific quality and methodological rigor       | 36  | 5.2  |
| Originality, authenticity, and contribution       | 31  | 4.5  |
| Integrity and transparency of AI use              | 161 | 23.5 |
| Traceability and evidence of the research process | 34  | 5.0  |
| Oral defence and justification of choices         | 53  | 7.7  |
| Rubrics, grids, and explicit criteria             | 55  | 8.0  |
| AI detection and evidential caution               | 40  | 5.8  |
| Final human decision and validation               | 4   | 0.6  |

Integrity and transparency of AI use were the most frequent jury-related dimension (23.5%), followed by explicit rubrics and criteria (8.0%), oral defence and justification of choices (7.7%), and AI detection/evidential caution (5.8%). These findings suggest that jury assessment should not rely on the manuscript alone but should integrate explicit criteria, oral justification, and selected evidence from the research process.

### Risks and limitations

**Table 7: Risks and limitations associated with GenAI in assessment**

| Risk / limitation                              | f   | %    |
|------------------------------------------------|-----|------|
| Integrity, plagiarism, and misconduct          | 143 | 20.8 |
| Hallucinations, errors, and reliability        | 58  | 8.5  |
| Dependence, substitution, and loss of autonomy | 40  | 5.8  |
| Bias, discrimination, and equity               | 49  | 7.1  |
| Data protection and confidentiality            | 6   | 0.9  |
| Uncertain detection and false positives        | 40  | 5.8  |
| Copyright and intellectual property            | 1   | 0.1  |
| Cognitive or information overload              | 6   | 0.9  |

The most frequent risk was integrity, plagiarism, and misconduct (20.8%). Hallucinations and reliability problems represented 8.5%, bias and equity 7.1%, dependence or loss of autonomy 5.8%, and uncertain detection/false positives 5.8%. This pattern confirms that detector-based approaches cannot, by themselves, secure valid assessment.<sup>28,25</sup>

### Assessment transformations and good practices

**Table 8: Assessment transformations and good practices proposed in the corpus**

| Transformation / practice               | f  | %   |
|-----------------------------------------|----|-----|
| Assessment redesign                     | 31 | 4.5 |
| Authentic and contextualized assessment | 41 | 6.0 |
| Transparency and declaration of AI use  | 39 | 5.7 |

| Transformation / practice              | f  | %    |
|----------------------------------------|----|------|
| AI literacy and stakeholder training   | 84 | 12.2 |
| Human validation/verification          | 31 | 4.5  |
| Rubrics and explicit criteria          | 28 | 4.1  |
| Process evidence and traceability      | 35 | 5.1  |
| Oral defence and evaluative dialogue   | 50 | 7.3  |
| Explicit task-level policies and rules | 98 | 14.3 |

Explicit task-level policies and rules were the most frequent transformation (14.3%), followed by AI literacy and training (12.2%). Oral defence and evaluative dialogue reached 7.3%, authentic assessment 6.0%, transparency 5.7%, and process evidence 5.1%. The literature is therefore moving from surveillance alone toward explicit governance, capability building, authentication, and traceability.

### Multiple Correspondence Analysis

**Table 9: Raw inertia of the first five MCA dimensions**

| Dimension | Raw inertia (%) | Cumulative (%) |
|-----------|-----------------|----------------|
| 1         | 8.05            | 8.05           |
| 2         | 7.63            | 15.67          |
| 3         | 6.07            | 21.74          |
| 4         | 5.65            | 27.39          |
| 5         | 4.78            | 32.16          |

The first three dimensions explained 21.74% of raw inertia. With 44 binary variables, such dispersion is expected; interpretation was therefore complemented by Benzécri and Greenacre corrections.

**Table 10: Benzécri- and Greenacre-corrected inertia**

| Axis | Benzécri (%) | Benzécri cumulative (%) | Greenacre (%) | Greenacre cumulative (%) |
|------|--------------|-------------------------|---------------|--------------------------|
| 1    | 30.64        | 30.64                   | 20.89         | 20.89                    |
| 2    | 26.33        | 56.97                   | 17.95         | 38.84                    |
| 3    | 13.23        | 70.20                   | 9.02          | 47.86                    |
| 4    | 10.46        | 80.66                   | 7.13          | 54.99                    |
| 5    | 5.76         | 86.42                   | 3.93          | 58.92                    |

After Benzécri correction, the first two dimensions represented 56.97% of corrected inertia and the first three 70.20%. Under Greenacre correction, the first three dimensions represented 47.86%. The structure therefore supports a primarily three-dimensional interpretation. Dimension 1 opposed a normative logic centred on integrity, plagiarism, authorship, and transparency to a criterion-based instrumentation logic involving rubrics and explicit criteria. Dimension 2 was structured by traceability, process evidence, and human validation. Dimension 3 was dominated by rubrics, authentic assessment, and oral or dialogic modalities. Together, the dimensions describe three poles: integrity safeguards; process verification; and criterion-based, dialogic authentication of competence.

### Phi associations

**Table 11: Selected Phi associations**

| Crossed variables                                                  | $\phi$ | $\chi^2$ | p      | Interpretation       |
|--------------------------------------------------------------------|--------|----------|--------|----------------------|
| Integrity-authorship-attribution × Integrity-plagiarism-misconduct | 0.906  | 563.15   | < .001 | Very strong positive |
| AI integrity/transparency × Integrity/plagiarism                   | 0.867  | 516.15   | < .001 | Very strong positive |
| Integrity-authorship × AI integrity/transparency                   | 0.842  | 486.45   | < .001 | Very strong positive |
| Design/scenario × Assessment redesign                              | 0.818  | 458.94   | < .001 | Very strong positive |

| Crossed variables                             | $\phi$ | $\chi^2$ | p      | Interpretation  |
|-----------------------------------------------|--------|----------|--------|-----------------|
| Equity/inclusion × Bias/discrimination/equity | 0.756  | 391.56   | < .001 | Strong positive |
| Oral/defence × Oral evaluative dialogue       | 0.738  | 373.59   | < .001 | Strong positive |
| Authentic assessment × Oral/defence           | 0.694  | 330.43   | < .001 | Strong positive |
| Rubrics/criteria × Automated grading          | 0.191  | 24.91    | < .001 | Weak positive   |
| Process traceability × Human validation       | 0.177  | 21.41    | < .001 | Weak positive   |
| Pedagogical feedback × Teacher practices/role | 0.144  | 14.17    | < .001 | Weak positive   |

Phi coefficients confirmed several proximities identified by MCA. The association between authentic assessment and oral/defence-based assessment was particularly strong ( $\phi = .694$ ;  $\chi^2(1) = 330.43$ ;  $p < .001$ ). Process traceability was also associated with human validation ( $\phi = .177$ ;  $\chi^2(1) = 21.41$ ;  $p < .001$ ), while rubrics/criteria were associated with automated grading ( $\phi = .191$ ;  $\chi^2(1) = 24.91$ ;  $p < .001$ ). Pedagogical feedback was weakly but significantly associated with teacher practices and role ( $\phi = .144$ ;  $\chi^2(1) = 14.17$ ;  $p < .001$ ). Very high coefficients between semantically close categories should be interpreted cautiously because coding proximity can inflate association.

Several theoretically plausible relations were not statistically established in the corpus: integrity/plagiarism × AI detection ( $\phi = .071$ ;  $p = .061$ ), integrity/plagiarism × transparency/declaration of AI use ( $\phi = .060$ ;  $p = .116$ ), traceability × oral/defence ( $\phi = .035$ ;  $p = .366$ ), AI literacy × explicit policies ( $\phi = .025$ ;  $p = .506$ ), hallucinations × human validation ( $\phi = -.016$ ;  $p = .682$ ), and bias/equity × human validation ( $\phi = -.033$ ;  $p = .386$ ). These null associations are important because they prevent conceptually plausible relationships from being presented as empirically demonstrated co-occurrences.

#### IV. DISCUSSION

The central finding of this study is that GenAI does not merely change assessment instruments; it challenges the nature of the evidence on which university judgment is based. Generative models can produce outputs capable of succeeding in some assessment settings, sometimes without detection, while automated grading and feedback can show useful but context-dependent agreement with human evaluation.<sup>22,7,4,17</sup> The implication is neither the abandonment of assessment nor the replacement of academic judgment by algorithms. It is the reconstruction of the conditions under which a performance can validly be attributed to the person being assessed.

##### A rapidly emerging scientific field

The concentration of 72.0% of documents in 2025–2026 shows how rapidly scholarship on GenAI and higher-education assessment has developed. This timing helps explain the methodological heterogeneity of the corpus and the prominence of conceptual analyses, reviews, and studies whose methods were incompletely specified. Recent reviews similarly describe a fast-expanding and still stabilizing field.<sup>1,3,8,19,27</sup>

This rapid evolution limits generalization. Recommendations derived from one model version, discipline, or institutional setting cannot automatically be transferred to all contexts. However, convergence across the corpus supports several broader principles: explicit rules, AI literacy, authentic tasks, process traceability, oralization when justified, and final human validation.

##### Assessing teaching: from tool control to pedagogical design quality

Within the teaching domain, teacher practices and roles (19.4%) and pedagogical feedback (18.7%) were much more frequent than the explicit phrase 'pedagogical quality'. This suggests that teaching quality in the GenAI era is being operationalized through what teachers design, regulate, explain, and verify. Evaluating teaching therefore increasingly requires examining the coherence among intended outcomes, assessment tasks, permitted AI assistance, feedback processes, and validation criteria.

The AIAS illustrates this shift by making the level of permitted AI assistance explicit and linking it to assessment design.<sup>18,9</sup> The broader institutional literature likewise shows movement from blanket prohibition toward differentiated policies and assessment redesign.

##### Assessing learning evidence: the weakening of the final product as sole proof

Examinations, tests, and quizzes remained the most frequent learning-assessment category (22.2%), followed by written work (16.0%). These are precisely the formats for which the literature has questioned whether generative models can produce plausible

## Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence

or high-performing responses.<sup>22,24,29</sup> The problem is not that these formats become useless, but that their evidential value depends more heavily on assessment conditions and complementary evidence.

Authentic assessment appeared in 8.6% of learning-assessment documents and was strongly associated with oral/defence-based assessment ( $\phi = .694$ ). This is one of the most substantively important findings. Within the corpus, authenticity is not limited to contextualized tasks; it is often paired with mechanisms that require the student to explain, defend, or reconstruct reasoning. This does not mean that every assessment should become oral. Rather, oral dialogue can serve as complementary evidence when the written product alone no longer establishes the origin or depth of individual reasoning.

### Research supervision: preserving authorship and scientific responsibility

In the supervision domain, integrity, authorship, and attribution represented 23.6% of documents, far exceeding the other categories. The key issue becomes one of intellectual responsibility: who formulated the problem, selected the method, verified the sources, interpreted the data, and assumed responsibility for the conclusions? Human-AI collaboration in academic writing makes this question especially salient.<sup>16,23</sup>

GenAI may assist reformulation, idea generation, structure, revision, and literature-related tasks, but assistance does not remove the researcher's duty to verify. Supervision therefore becomes more process-oriented. Intermediate versions, decision logs, analysis matrices, source trails, methodological notes, and declarations of relevant AI use can function as formative evidence. The aim is not continuous surveillance, but visibility at critical points where scientific responsibility must remain attributable to the researcher.

### Academic juries: from the isolated manuscript to triangulated evidence

For the jury function, integrity and transparency of AI use were dominant (23.5%), followed by explicit rubrics/criteria (8.0%) and oral defence/justification (7.7%). A jury can no longer infer scientific mastery solely from the polished manuscript when writing, coding, analysis, or editing may have been substantially assisted by GenAI. The defence consequently acquires an epistemic function: it allows the candidate to justify problem construction, source selection, design, analytical operations, interpretation, limitations, and contribution.

The findings do not justify turning automated detection into decisive evidence. The corpus explicitly identifies uncertain detection and false positives, and empirical detector evaluations have shown substantial reliability limitations.<sup>28</sup> Jury assessment should therefore prioritize converging evidence rather than infer misconduct from a detector score alone.

### Risks extend beyond plagiarism

Integrity, plagiarism, and misconduct were the most frequent risks (20.8%), but they did not exhaust the problem. Hallucinations and errors (8.5%), bias and equity (7.1%), dependence and loss of autonomy (5.8%), and uncertain detection/false positives (5.8%) show that responsible assessment cannot be reduced to fraud prevention. A student may transparently declare AI use and still reproduce false information; conversely, human-authored work may be falsely classified as AI-generated. Responsible evaluation therefore requires transparency, content verification, explicit criteria, and an opportunity for the author to justify decisions.

### Interpreting MCA and Phi together

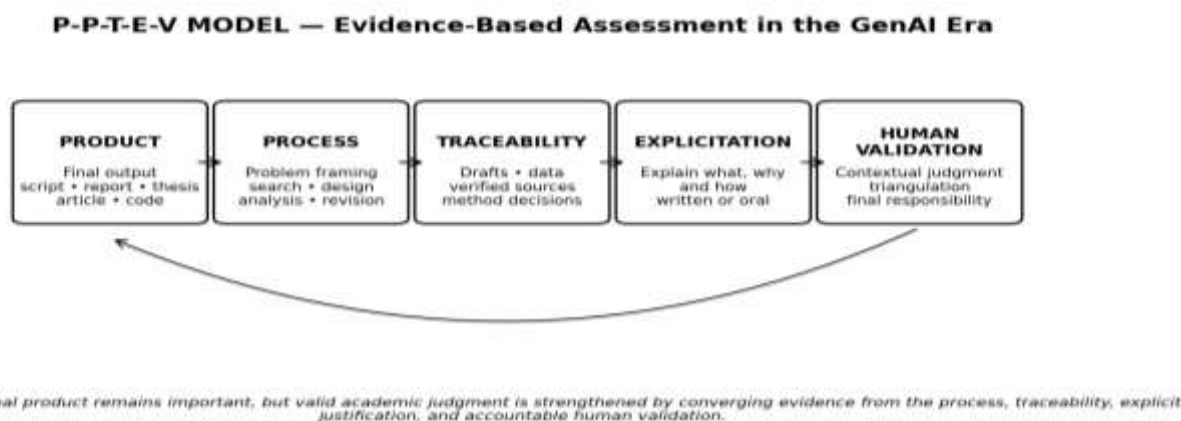
The MCA supports three complementary assessment logics. The first concerns integrity, authorship, transparency, and explicit criteria: a normative and criterion-based safeguarding function. The second concerns traceability and human validation: a verifiability function. The third links rubrics, authenticity, and oral/dialogic modalities: an authentication function. These are not competing models but complementary functions of the same assessment system—establish the rules, document the process, and verify mastery through contextualized and dialogic evidence.

Phi coefficients help distinguish empirical co-occurrence from conceptual proximity. Very high coefficients between semantically overlapping integrity categories require caution. More informative are associations linking distinct dimensions, particularly authentic assessment with oral/defence-based assessment ( $\phi = .694$ ), process traceability with human validation ( $\phi = .177$ ), and rubrics/criteria with automated grading ( $\phi = .191$ ). The non-significant associations are equally useful because they show that theoretical coherence should not be confused with empirical co-occurrence. Given the 946 pairwise tests, confirmatory use should include false-discovery-rate control.

### The P-P-T-E-V integrative model

Based on the convergence of descriptive frequencies, MCA, and Phi associations, this study proposes the P-P-T-E-V model: Product, Process, Traceability, Explication, and Human Validation. It is an interpretive model derived from the corpus and requires empirical testing.

Figure 1: P-P-T-E-V model for evidence-based assessment in the GenAI era



Product refers to the final object of assessment—script, report, thesis, article, project, code, or presentation. It remains indispensable but is no longer treated as self-sufficient evidence of individual mastery. Process refers to the intellectual and methodological steps that produced the output: problem formulation, literature search, design, analysis, revision, and decision-making. Traceability consists of selected evidence of that process, such as intermediate versions, data, analysis matrices, verified references, methodological decisions, or declarations of relevant GenAI use. Explicitation is the ability of the student or researcher to explain what was done, why it was done, how AI was used when relevant, and how outputs were verified. Human validation means that final academic responsibility remains with a lecturer, supervisor, examiner, or jury capable of triangulating evidence and exercising contextualized judgment.

The model is dynamic rather than linear. The product raises questions about the process; the process generates traces; traces support explicit justification; explicitation informs human validation; and validation may lead the evaluator to revisit the product or specific stages of the process. The objective is not to monitor every action but to obtain a sufficient bundle of converging evidence to support a valid academic decision.

## Practical implications

For teachers, the findings support specifying for each task whether GenAI use is prohibited, permitted, or expected; aligning this rule with the intended learning outcome; using explicit criteria; requesting process evidence or brief oral justification when stakes warrant it; and training students to verify generated outputs. Assessment redesign does not necessarily require entirely new tasks; it requires stronger alignment between what is to be learned and what counts as acceptable evidence of that learning.

For thesis and dissertation supervisors, acceptable AI uses, disclosure requirements, and stages requiring personal validation should be defined from the beginning. Supervision meetings can focus on scientific decisions rather than only textual correction. Important sources, data, analyses, and interpretations should remain explainable and verifiable by the researcher.

For academic juries, four evidence sources can be triangulated: manuscript quality, methodological coherence, sufficient process traceability, and the candidate's ability to justify choices. Defence rubrics can explicitly distinguish subject knowledge, methodological mastery, interpretation, contribution, transparency of AI use, and capacity to answer questions. AI detectors, if used at all, should not constitute the sole basis of a high-stakes decision.

For institutions, the frequency of explicit task-level policies (14.3%) and AI literacy (12.2%) indicates that the response cannot be purely disciplinary. Universities need clear frameworks, staff and student training, discipline-sensitive guidance, and procedures ensuring that high-stakes decisions remain subject to meaningful human review.

## Scientific contribution, strengths, and limitations

The study contributes an integrated reading of 686 documents by connecting dimensions often examined separately: teaching, learning evidence, research supervision, jury assessment, risks, and assessment transformations. MCA indicates that these contents organize around three complementary logics—integrity and criteria; traceability and validation; authenticity and dialogue—while Phi coefficients distinguish observed associations from merely theoretical links. The main conceptual contribution is the shift from 'AI detection' to 'validity of assessment evidence', formalized through the P-P-T-E-V model.

Three strengths are notable: the size of the corpus; the combination of descriptive analysis, MCA, corrected inertia, and Phi coefficients; and the explicit distinction between research supervision and jury assessment. Several limitations must also be recognized. First, thematic coding was systematic and lexical-thematic but was not independently double-coded with an inter-coder agreement statistic. Second, some categories are semantically close, which can inflate Phi coefficients; exact duplicate pairs were excluded from interpretation. Third, the corpus is heterogeneous across disciplines, methods, populations, and institutional contexts. Frequencies describe the literature collected, not the prevalence of practices in all universities. Fourth, jury-related categories include

## Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence

principles transferable to jury assessment and are not limited to empirical studies of thesis or dissertation juries. Finally, because 946 Phi tests were calculated, false-discovery-rate correction is necessary for exhaustive confirmatory inference.

### Future research

Future studies should first manually validate the coding system on a representative subsample and calculate inter-coder agreement. MCA should then be repeated after merging redundant categories to test the stability of the dimensions. Empirical studies should test the P-P-T-E-V model with teachers, students, supervisors, and jury members. Quasi-experimental designs could compare product-only assessment with assessment combining product, process evidence, and oral justification. Outcomes should include validity of judgment, inter-rater agreement, assessment time, acceptability, equity, and the ability to identify actual mastery.

A further priority concerns doctoral research and master's theses. Specific rubrics should be developed and validated to assess transparency of GenAI use without conflating legitimate assistance with misconduct. Cross-disciplinary and cross-national comparisons would also help identify which components of the model are broadly transferable and which require contextual adaptation.

## V. CONCLUSION

This study analysed 686 documents published between 2022 and 2026 to examine transformations in the assessment of teaching, learning evidence, and scientific research in higher education in the era of GenAI. The corpus, overwhelmingly recent, shows that integrity, transparency, explicit policies, AI literacy, authenticity, traceability, oral justification, and human validation are recurrent dimensions of the emerging assessment landscape.

MCA identified three principal structures: integrity and objectification of criteria; traceability and validation; and authentic, criterion-based, dialogic assessment. After Benzécri correction, the first three dimensions represented 70.20% of corrected inertia; under Greenacre correction, they represented 47.86%. Phi coefficients further showed a strong association between authentic assessment and oral/defence-based assessment ( $\phi = .694$ ) and a more modest positive association between traceability and human validation ( $\phi = .177$ ).

The central conclusion is that GenAI makes the binary choice between authorization and prohibition less important than the question of what constitutes valid evidence of learning or scientific mastery. Assessment can no longer always rely on the final product alone. It is strengthened when product, process, traceability, explicitation, and human validation are combined. The P-P-T-E-V model formalizes this shift and provides a framework for future empirical testing.

Higher education therefore retains its full evaluative function, but academic judgment becomes more explicitly grounded in the convergence of evidence. The goal is not to make all technological assistance impossible; it is to preserve the capacity of assessment to establish what the student or researcher understands, can do, can explain, and can assume intellectual responsibility for.

## VI. ACKNOWLEDGMENTS

The author acknowledges the scholarly contributions of the researchers whose publications formed the documentary corpus analysed in this study. No individual who does not meet authorship criteria is presented as an author.

## VII. DISCLOSURE

Funding: This research received no specific grant from any public, commercial, or not-for-profit funding agency and was conducted without external financial support. Conflict of interest: The author declares no conflict of interest that could have influenced the conduct, analysis, interpretation, or presentation of this study.

## REFERENCES

1. Batista, J., Mesquita, A., & Carnaz, G. (2024). Generative AI and higher education: Trends, challenges, and future directions from a systematic literature review. *Information*, 15(11), 676. <https://doi.org/10.3390/info15110676>
2. Bearman, M., Ryan, J., & Ajjaw, R. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher Education*. <https://doi.org/10.1007/s10734-024-01265-3>
3. Belkina, M., Daniel, S., Nikolic, S., Haque, R., Lyden, S., Neal, P., Grundy, S., & Hassan, G. M. (2025). Implementing generative AI (GenAI) in higher education: A systematic review of case studies. *Computers and Education: Artificial Intelligence*, 100407. <https://doi.org/10.1016/j.caeai.2025.100407>
4. Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7, 100299. <https://doi.org/10.1016/j.caeai.2024.100299>
5. El-Adawy, S., MacDonagh, A., & Abdelhafez, M. (2024). Exploring large language models as formative feedback tools in physics. *2024 Physics Education Research Conference Proceedings*, 126–131. <https://doi.org/10.1119/perc.2024.pr.El-Adawy>

6. Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, 57. <https://doi.org/10.1186/s41239-023-00425-2>
7. Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, 51(1), 201–224. <https://doi.org/10.1002/berj.4069>
8. Fu, Y., Weng, Z., & Wang, J. (2025). Examining AI use in educational contexts: A scoping meta-review and bibliometric analysis. *International Journal of Artificial Intelligence in Education*, 35, 1388–1444. <https://doi.org/10.1007/s40593-024-00442-w>
9. Furze, L., Perkins, M., Roe, J., & MacVaugh, J. (2024). The AI Assessment Scale (AIAS) in action: A pilot implementation of GenAI-supported assessment. *Australasian Journal of Educational Technology*, 40(4). <https://doi.org/10.14742/ajet.9434>
10. Lachheb, A., Leung, J., Abramenska-Lachheb, V., & Sankaranarayanan, R. (2025). AI in higher education: A bibliometric analysis, synthesis, and a critique of research. *The Internet and Higher Education*, 67, 101021. <https://doi.org/10.1016/j.iheduc.2025.101021>
11. Lodge, J., Howard, S., Bearman, M., Dawson, P., & Associates. (2023). Assessment reform for the age of artificial intelligence. Tertiary Education Quality and Standards Agency.
12. Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the ‘originality’ of students’ work. *Assessment & Evaluation in Higher Education*, 49(5), 651–664. <https://doi.org/10.1080/02602938.2024.2309963>
13. Lye, C. Y., & Lim, L. (2024). Generative artificial intelligence in tertiary education: Assessment redesign principles and considerations. *Education Sciences*, 14(6), 569. <https://doi.org/10.3390/educsci14060569>
14. Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, 5, 100151. <https://doi.org/10.1016/j.caeo.2023.100151>
15. Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or human? Evaluating student feedback perceptions in higher education. *Artificial Intelligence in Education*. <https://doi.org/10.31219/osf.io/6zm83>
16. Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, 49, 847–864. <https://doi.org/10.1080/03075079.2024.2323593>
17. Pecuchova, J., Benko, L., & Drlik, M. (2025). Automated grading of open-ended questions in higher education using GenAI models. *International Journal of Artificial Intelligence in Education*, 35(6), 3813–3846. <https://doi.org/10.1007/s40593-025-00517-2>
18. Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching & Learning Practice*, 21(6). <https://doi.org/10.53761/q3azde36>
19. Qian, Y. (2025). Pedagogical applications of generative AI in higher education: A systematic review of the field. *TechTrends*, 69, 1105–1120. <https://doi.org/10.1007/s11528-025-01100-1>
20. Rubin, M. (2023). Use and misuse of corrections for multiple testing. *Methods in Psychology*, 8, 100120. <https://doi.org/10.1016/j.metip.2023.100120>
21. Sahar, R., & Munawaroh, M. (2025). Artificial intelligence in higher education with bibliometric and content analysis for future research agenda. *Discover Sustainability*, 6, 401. <https://doi.org/10.1007/s43621-025-01086-z>
22. Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A ‘Turing Test’ case study. *PLOS ONE*, 19(6), e0305354. <https://doi.org/10.1371/journal.pone.0305354>
23. Song, N. (2024). Higher education crisis: Academic misconduct with generative AI. *Journal of Contingencies and Crisis Management*. <https://doi.org/10.1111/1468-5973.12532>
24. Soulage, C. O., Van Coppenolle, F., & Guebre-Egziabher, F. (2024). The conversational AI ‘ChatGPT’ outperforms medical students on a physiology university examination. *Advances in Physiology Education*, 48(4). <https://doi.org/10.1152/advan.00181.2023>
25. Susnjak, T., & McIntosh, T. R. (2024). ChatGPT: The end of online exam integrity? *Education Sciences*, 14(6), 656. <https://doi.org/10.3390/educsci14060656>
26. Ma, T. (2024). Systematically visualizing ChatGPT used in higher education: Publication trend, disciplinary domains, research themes, adoption and acceptance. *Computers and Education: Artificial Intelligence*, 100336. <https://doi.org/10.1016/j.caeai.2024.100336>
27. Wu, F., Dang, Y., & Li, M. (2025). A systematic review of responses, attitudes, and utilization behaviors on generative AI for teaching and learning in higher education. *Behavioral Sciences*, 15(4), 467. <https://doi.org/10.3390/bs15040467>

## Assessing Teaching and Scientific Research in Higher Education in the Era of Generative Artificial Intelligence

28. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, 26. <https://doi.org/10.1007/s40979-023-00146-z>
29. Wood, D. A., Achhpilia, M. P., Adams, M. T., Aghazadeh, S., Akinyele, K., Akpan, M., et al. (2023). The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education*, 38(4), 81–108. <https://doi.org/10.2308/ISSUES-2023-013>
30. Xia, Q., Weng, X., Ouyang, F., Lin, T. J., & Chiu, T. K. F. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, 21, 40. <https://doi.org/10.1186/s41239-024-00468-z>