



# INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH AND INNOVATIVE STUDIES

Advancing Knowledge, Inspiring Innovation

www.ijsrisjournal.com



**PEER REVIEWED**  
Rigorous & Transparent

**DOI: CROSSREF**  
Member

**OPEN ACCESS**  
For Global Visibility

**BI-MONTHLY**  
Timely Publication

**WIDE READERSHIP**  
High Impact & Visibility

**AUTHOR FRIENDLY**  
Fast | Supportive | Reliable

**RECEIVED**  
17 August 2026

**ACCEPTED**  
09 September 2026

**PUBLISHED**  
1 October 2026

**October 2026**

**Volume 5**

**Number 5**

**437 to 450**

## Multiple-choice questions as a means of assessment in the face of generative AI: a Python analysis of master's students' perceptions

Ahmed M'HAMDI<sup>1</sup>

<sup>1</sup>Higher School of Education and Training, Ibnou Zohr University, Morocco, [a.mhamdi@uiz.ac.ma](mailto:a.mhamdi@uiz.ac.ma)

### LICENSE



This article is licensed under the  
Creative Commons Attribution  
4.0 International License (CC BY 4.0).

https://creativecommons.org/licenses/by/4.0/



PREFIX :

10.63883

IJSRIS JOURNAL

### OPEN ACCESS



This is an open access article  
distributed under the terms of  
open access.

## ABSTRACT

The justifiable multiple-choice question, which requires students to select an answer and then justify it, is often presented as a defence against the unauthorised use of generative AI. This case study focuses on 41 of the 50 Master's students at the Higher School of Education and Training (Ibnou Zohr University) who sat an exam in this format (response rate: 82 per cent). Data from a mixed-

methods questionnaire were processed using a reproducible analysis pipeline in Python: descriptive statistics, reliability, non-parametric tests, k-means clustering and automated analysis of open-ended responses. The students recognise the format's high educational value (71 per cent believe it forces them to understand better;  $\alpha = 0.85$ ), but only 25 per cent consider it more difficult to circumvent using AI (Wilcoxon,  $p < 0.001$ ,  $r = 0.62$ ). This scepticism is shared by both identified profiles and does not depend on the frequency of AI use. Students attribute the resistance to the contextual nature

of the questions (7 per cent consider a question relating to classroom experience to be vulnerable) and the majority suggest redesigning the tasks rather than increasing supervision (71 per cent versus 22 per cent).

**Keywords:** Justification-based multiple-choice questions; generative AI; academic integrity; authentic assessment; educational data analysis; Python; case study.

## 1. Introduction

Since ChatGPT was made publicly available in November 2022, generative artificial intelligence (AI) tools have become capable of answering, within seconds, a large proportion of university examination questions. Multimodal versions exacerbate this phenomenon: a question photographed with a mobile phone is now sufficient to obtain a written answer. Several studies view this as a direct threat to the integrity of assessment (Cotton et al., 2024; Rudolph et al., 2023; Susnjak and McIntosh, 2024).

Faced with this reality, institutions are torn between two strategies. The first involves stepping up supervision and a return to paper-based exams. The second involves rethinking the design of assessment tasks so that they evaluate what AI struggles to replicate: contextual reasoning, personal experience and reasoned justification (Lodge et al., 2023; Moorhouse et al., 2023; Perkins et al., 2024).

The ‘justified multiple-choice’ question lies at the intersection of these two approaches. Students select an option and then justify their choice in a few lines. This format, which was already studied in the 1980s (Nield and Wintre, 1986; Dodd and Leal, 1988), retains the effectiveness of multiple-choice questions whilst requiring a record of the reasoning process. It is often presented as being more resistant to cheating and AI. However, this hypothesis remains poorly documented from the students’ own perspective.

This research focuses on the examination for the module ‘New Technologies and AI in Science Education’, scheduled for the first semester of the Master’s degree in Educational Engineering and Science Didactics at the Higher School of Education and Training (ESEF), Ibn Zohr University. This examination was taken in the ‘multiple-choice with justification’ format by students on the Master’s programme. These students occupy a dual role: whilst they are being assessed today, tomorrow they will be teachers or trainers – and thus designers of assessments.

The study addresses four research questions:

1. What pedagogical value do students attribute to the multiple-choice question (MCQ) with justifications?
2. Do they consider them resistant to generative AI, and which examination formats do they perceive as vulnerable?
3. What student profiles emerge from their responses?
4. What suggestions do they offer for designing assessments that are both fair and difficult for AI to circumvent?

Methodologically, the article presents an analysis workflow implemented entirely in Python and designed to be reproducible. It combines descriptive statistics, non-parametric tests, unsupervised classification and automated analysis of textual responses.

## 2. Theoretical framework

### 2.1. From traditional multiple-choice questions to open-ended multiple-choice questions

Multiple-choice questions offer rapid, objective marking and broad coverage of the syllabus. Their limitations are well known: passing by chance, mere recognition of the correct answer and the encouragement of superficial learning (Scouller, 1998; Burton, 2005). Justified multiple-choice questions address these limitations by asking students to explain their choice. Justification makes the reasoning process visible, reduces the effect of chance and enables the teacher to distinguish between a correct answer that has been understood and one that has been guessed (Nield and Wintre, 1986; Dodd and Leal, 1988).

### 2.2. Authentic assessment and pedagogical alignment

Authentic assessment refers to tasks that closely resemble real-life situations in which the relevant skills will be applied (Wiggins, 1990). Gulikers et al. (2004) describe five dimensions of this: the task, the physical environment, the social context, the expected outcome and the criteria. The principle of constructive alignment (Biggs, 1996) also emphasises that the form of assessment shapes the way in which students learn. An assessment that requires reasoning promotes deep learning.

### 2.3. Generative AI and the integrity of assessment

Large language models open up real pedagogical possibilities but undermine assessments based on reproduction (Kasneci et al., 2023). Susnjak and McIntosh (2024) show that they produce high-quality answers in remote examinations, including for reasoning questions. Several frameworks now propose grading the permitted use of AI rather than banning it, such as the AIAS scale by Perkins et al. (2024). The issue is therefore no longer simply one of detecting AI, but of designing tasks where its undeclared use adds little value.

### 2.4. Computational analysis of educational data

Educational data mining and learning analytics utilise computational methods to gain an understanding of learners based on their digital footprints (Romero and Ventura, 2010; Siemens and Baker, 2012). These methods are not limited to large- -volume datasets: even on a small corpus, unsupervised classification and automated text analysis help to structure the exploration and make the process reproducible. The Python ecosystem (pandas, SciPy, scikit-learn) offers open-source and well-documented tools for this purpose (McKinney, 2010; Pedregosa et al., 2011).

## 3. Methodology

### 3.1. Research design: an instrumental case study

The study adopts the instrumental case study design (Yin, 2018): the case is studied not for its own sake, but to shed light on a broader question, namely the role of the justified multiple-choice questionnaire in assessment in the era of generative AI. The approach is mixed, with a quantitative focus. It combines closed-ended data (Likert scales, multiple-choice questions) and textual data (open-ended questions), processed within a single analysis pipeline in Python.

The case is defined as follows: all ESEF Master's students who sat the examination for the module 'New Technologies and AI in Science Education' during the 2025–2026 academic year, totalling 50 students. This choice is driven by a desire for in-depth understanding rather than statistical generalisation: the validity of the study rests on the coverage of the case and on the triangulation of the analyses, not on the representativeness of a sample.

The exam took place in February 2026, at the end of the first semester. Lasting 2 hours 30 minutes, it was held in person, on the same subject and under the same conditions for all students. Each question offered several options; students had to select the correct answer and then justify their choice in writing, with no imposed length limit. The marking took both

components into account: a quarter of the marks for each question related to the choice of option and three-quarters to the justification, so that a correct answer that was not justified or poorly justified earned only a portion of the marks. The exam was designed and marked by the module tutor, who is also the author of this study.

### 3.2. Participants

The questionnaire received 48 responses. Seven were from undergraduate students (mathematics) who were not part of the study; these were excluded prior to any analysis. The final sample comprised 41 Master's students out of the 50 in the study, representing a response rate of 82 per cent. The inclusion criteria were twofold: being enrolled on the Master's programme and having sat the module examination in the multiple-choice with justification format.

**Table 1:** Profile of respondents.

| Characteristic                  | Mode                | <i>n</i> | %    |
|---------------------------------|---------------------|----------|------|
| Gender                          | Female              | 28       | 68.3 |
|                                 | Men                 | 13       | 31.7 |
| Age                             | 20 to 22 years      | 14       | 34.1 |
|                                 | 23 to 25 years      | 3        | 7.3  |
|                                 | Over 25             | 24       | 58.5 |
| Use of generative AI in studies | Very often or daily | 10       | 24.4 |
|                                 | Often               | 21       | 51.2 |
|                                 | Sometimes           | 9        | 22.0 |
|                                 | Rarely              | 1        | 2.4  |
| Proficiency with digital tools  | Very good           | 5        | 12.2 |
|                                 | Good                | 27       | 65.9 |

|  |         |   |      |
|--|---------|---|------|
|  | Average | 7 | 17.1 |
|  | Low     | 2 | 4.9  |

*Note: n = 41.*

The group is predominantly female (68 per cent) and on the older side: 59 per cent of respondents are over 25. The use of generative AI is widespread: three out of four students use it frequently or daily in their studies, and 78 per cent rate their digital literacy as good or very good. The sample therefore comprises users who are familiar with these tools, which lends weight to their assessment of how resistant exams are to AI.

### 3.3. Data collection instrument

The instrument is an online questionnaire designed for the study and administered via Google Forms. It comprises 28 questions organised into six sections (Table 2). The Likert scales consist of five points, ranging from 1 (strongly disagree) to 5 (strongly agree), with a neutral midpoint.

**Table 2:** Questionnaire structure.

| Block                                                                  | Questions | Response format                                  | Content                                                                                                   |
|------------------------------------------------------------------------|-----------|--------------------------------------------------|-----------------------------------------------------------------------------------------------------------|
| A. Profile                                                             | Q1 to Q5  | Single choice                                    | Educational background, gender, age group, frequency of use of generative AI, digital literacy            |
| B. Perception of the multiple-choice questionnaire with justifications | Q6 to Q14 | 5-point Likert scale                             | Clarity, comprehensibility, validity, time required, randomness, copying, AI, usefulness, overall opinion |
| C. Advantages and disadvantages                                        | Q15, Q16  | Single choice (with open-ended response for Q15) | Perceived main advantage and main disadvantage                                                            |
| D. Examination formats                                                 | Q17, Q18  | Multiple-choice (8 formats); single-choice       | Formats considered suitable for assessing skills; need to modify the assessment                           |

|                                 |                 |                                                   |                                                                                                  |
|---------------------------------|-----------------|---------------------------------------------------|--------------------------------------------------------------------------------------------------|
| E. Generative AI and assessment | Q19 to Q25; Q26 | 5-point Likert scale; multiple-choice (6 formats) | Views on AI and assessment; formats that AI could easily solve using a photograph of the subject |
| F. Open-ended questions         | Q27, Q28        | Free-form text                                    | Suggestions for designing fair exams that are difficult to cheat on; other comments              |

The sixteen Likert-scale items are reproduced in full in Table 3, together with the code used in the analyses. Item 9 is worded to measure a perceived cost (time); it is therefore not reversed but interpreted separately. Item 12 specifically concerns the format's resistance to AI, whilst item 22 concerns justification in general: this pair makes it possible to distinguish between judgements on the principle of justification and judgements on the specific format.

**Table 3:** Likert-scale items in the questionnaire.

| Code | Item wording                                                                                                                |
|------|-----------------------------------------------------------------------------------------------------------------------------|
| Q6   | The instructions for this type of exam (multiple-choice questions + justification) were clear.                              |
| Q7   | Having to justify my answers forced me to understand the content better, rather than simply identifying the correct answer. |
| Q8   | This format better reflects my actual knowledge than a standard multiple-choice test without explanations.                  |
| Q9   | The requirement to provide justifications made the exam more time-consuming.                                                |
| Q10  | This format reduces the chance of passing by sheer luck or by guessing.                                                     |
| Q11  | This format makes it more difficult for students to cheat or help one another without authorisation.                        |
| Q12  | This type of exam makes it more difficult to use generative AI to answer on my behalf.                                      |
| Q13  | I found the effort involved in justifying my answers useful for my learning.                                                |

|     |                                                                                                |
|-----|------------------------------------------------------------------------------------------------|
| Q14 | Overall, I have a positive view of this type of exam.                                          |
| Q19 | Generative AI makes it easier to cheat in exams and homework.                                  |
| Q20 | Face-to-face exams (without connected devices) are more reliable in the face of generative AI. |
| Q21 | Oral exams are a better way of assessing what a student has actually understood.               |
| Q22 | Asking for justification or reasoning makes the assessment more resistant to generative AI.    |
| Q23 | Generative AI can be a useful learning tool if its use is properly supervised.                 |
| Q24 | I would be in favour of examinations where the use of AI is permitted but explicitly stated.   |
| Q25 | Teachers should adapt their exam questions so that they are more difficult for AI to handle.   |

*Note:* Five-point scale: 1 = strongly disagree; 5 = strongly agree.

Question 26 warrants further clarification, as it lies at the heart of the study. It places the student in a concrete scenario: ‘Would a photographed exam paper submitted to a generative AI easily produce the correct answer?’. They then tick, from a choice of six formats (standard multiple-choice, multiple-choice with justification, questions on course content or definitions, contextualised case studies, questions relating to a real-life experience in class or during practical sessions, diagrams or figures to analyse), those for which the AI would provide the answer easily. This situational, rather than abstract, wording aims to elicit a judgement grounded in students’ actual experiences.

### 3.4. Data collection procedure and ethical considerations

The questionnaire was distributed by the module tutor via WhatsApp, approximately four months after the exam, once the results were known. Responses from Master’s students were collected between 10 June and 7 July 2026. Participation was voluntary and unrelated to the module assessment. The first question sought informed consent: by ticking the box, the student agreed that their anonymous responses could be used for research purposes, and only responses accompanied by

this consent were retained (48 out of 48). The form did not collect names, email addresses or student numbers. Prior to analysis, the timestamps were removed and each respondent was assigned a neutral identifier (E01 to E48, allocated in the order in which responses were received, before case filtering). The extracts from open-ended responses cited in the article are identified solely by these identifiers.

### 3.5. Data preparation

The raw export from Google Forms (a UTF-8-encoded CSV file, 48 rows and 30 columns) was processed using the pandas library. The preparation involved six steps, all of which are documented in the notebook.

1. **Renaming.** Long question titles were replaced with short codes (Q6 to Q25 for Likert items, descriptive names for the other variables).
2. **Anonymisation and filtering.** The timestamp and consent column were removed, identifiers were assigned, and only Master’s students were retained ( $n = 41$ ).
3. **Recoding.** Likert items are retained as ordinal variables ranging from 1 to 5. For descriptive analyses, they are grouped into three categories: disagreement (1–2), neutral (3) and agreement (4–5). The frequency of AI use is recoded into an ordinal scale ranging from 1 (rarely) to 4 (very often or daily).
4. **Splitting of multiple-choice answers.** Responses to questions 17 and 26, stored as a string separated by semicolons, are broken down into binary variables (one per format, 1 = ticked).
5. **Closed-ended free-text responses.** For question 15, four respondents entered a free-text response instead of selecting a suggested option; these responses are retained as they are and are not reclassified into a category.
6. **Missing data.** Only one respondent (E42) has missing values on Likert-scale items (items 10 to 14, i.e. 5 values out of 656, less than 1 per cent). Two respondents did not answer question 16, and one did not answer question 15. Descriptive analyses and tests use all available responses for each item or pair of items (pairwise deletion). For classification purposes only, which requires a complete matrix, missing values are replaced by the median for that item.

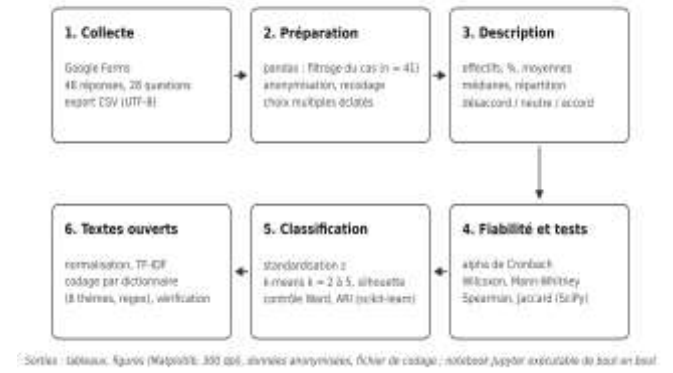
Finally, an educational value score was calculated for each student, as the average of their responses to the seven items retained following the reliability analysis (items 6, 7, 8, 10, 11, 13 and 14). For the cross-typology in section 4.2, a score of 3.5 or higher is considered high, as this threshold corresponds to the midpoint of the agreement zone.

### 3.6. Analysis framework

Table 4 links each research question to the data used and the methods employed. Figure 1 presents the complete data processing workflow.

**Table 4:** Analysis plan by research question.

| Research question                                                   | Data                                       | Methods                                                                                               | Python tools                  |
|---------------------------------------------------------------------|--------------------------------------------|-------------------------------------------------------------------------------------------------------|-------------------------------|
| <b>QR1. Perceived educational value</b>                             | Items 6 to 14;<br>questions 15 and 16      | Descriptive statistics;<br>Cronbach's alpha;<br>item-total correlations;<br>Spearman's correlations   | pandas,<br>SciPy              |
| <b>QR2. Perceived resistance to AI and vulnerability of formats</b> | Items 12, 19 to 25;<br>questions 17 and 26 | Paired Wilcoxon test; Spearman;<br>Jaccard index                                                      | SciPy,<br>pandas              |
| <b>QR3. Student profiles</b>                                        | Items 6 to 14, 22, 23 and 24               | Standardisation; k-means; silhouette index; Ward's clustering; adjusted Rand index; Mann-Whitney test | scikit-learn                  |
| <b>QR4. Proposed design approaches</b>                              | Questions 27 and 28                        | Text normalisation; TF-IDF; dictionary-based thematic coding; echo index                              | scikit-learn, re, unicodedata |



**Figure 1:** Python analysis pipeline, from data collection to results.

*Note:* Each step corresponds to a section of the Jupyter notebook provided with the article.

### 3.7. Statistical analyses

**Descriptive statistics.** For each Likert item, we calculate the frequency, mean, standard deviation, median and the distribution across ‘disagree’, ‘neutral’ and ‘agree’ responses. As the Likert data are ordinal, medians and percentages of agreement are prioritised in the interpretation; means are reported to facilitate comparisons.

**Reliability.** The internal consistency of the nine items in Block B is estimated using Cronbach’s alpha (Cronbach, 1951), where  $k$  is the number of items,  $\sigma_i^2$  is the variance of item  $i$  and  $\sigma_T^2$  is the variance of the total score:

$$\alpha = \frac{k}{k-1} \left[ 1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_T^2} \right]$$

For each item, we also calculate the Spearman’s correlation between the item and the sum of the other items (item-rest correlation) and the alpha obtained if the item is removed. An item with a low item-rest correlation and whose removal increases the alpha does not measure the same dimension as the others.

**Paired comparisons.** To determine whether the same student considers the format to be more formative than AI-resistant, the responses to two items are compared using the Wilcoxon signed-rank test (two-tailed; null differences are excluded, using the ‘zero\_method = “wilcox”’ option in SciPy). The effect size is calculated from the Z-statistic and the number of pairs  $N$ :

$$r = \frac{|Z|}{\sqrt{N}}$$

The usual thresholds are 0.10 (small effect), 0.30 (medium effect) and 0.50 (large effect).

**Associations.** The relationships between ordinal items are measured using Spearman's correlation coefficient ( $\rho$ ). The full matrix is provided in the appendix (Figure A1).

**Comparison of independent groups.** The two profiles resulting from the classification (section 3.8) are compared using the Mann-Whitney test (two-tailed), with the effect size being the biserial rank correlation  $r_{rb}$ , where  $U$  is the test statistic and  $n_1, n_2$  are the sample sizes of the two groups:

$$r_{rb} = 1 - \frac{2U}{n_1 \times n_2}$$

**Associations between formats.** For the formats ticked in question 17, the similarity between two formats A and B is measured by the Jaccard index, i.e. the proportion of students who ticked both formats amongst those who ticked at least one of the two:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The significance threshold is set at 0.05. No correction for multiple comparisons is applied: comparisons between profiles are explicitly presented as exploratory, and the interpretation is based primarily on effect sizes.

### 3.8. Unsupervised classification

The classification aims to ascertain whether subgroups of students differ in their attitudes towards the multiple-choice questionnaire and AI, and in particular whether scepticism towards AI is specific to a particular subgroup. It focuses on twelve variables: the nine items in Block B and items 22, 23 and 24, which describe attitudes towards AI in assessment. Each variable is standardised ( $z = (x - \mu) / \sigma$ ) so that they all carry equal weight in the calculation of distances.

The k-means algorithm (scikit-learn implementation) is run for a number of clusters  $k$  ranging from 2 to 5, with 50 runs per value of  $k$  and a fixed random seed (`random_state = 42`) to ensure reproducibility. The number of clusters is chosen based on the average silhouette index (Rousseeuw, 1987). For each student  $i$ ,  $a(i)$  is the average distance to the members of their cluster and  $b(i)$  is the average distance to the nearest neighbouring cluster:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

The index ranges from  $-1$  to  $1$ ; the higher it is, the better the groups are separated. The robustness of the selected solution is assessed in two ways. Firstly, an ascending hierarchical clustering using Ward's method (Ward, 1963) is applied to the same data, and its agreement with k-means is measured by the adjusted Rand index (ARI, equal to 1 for two identical partitions and close to 0 for agreement due to chance). Next, k-means is run 200 times with a single initialisation and a different seed, and the average ARI with the reference solution measures stability. The resulting groups are named retrospectively according to their average educational value score, then compared on item 12 using a Mann-Whitney test.

### 3.9. Automated analysis of open-ended responses

**Corpus.** For each student, the answers to questions 27 and 28 are combined into a single text, forming a corpus of 41 documents. Responses lacking content ('no', 'nothing', 'no idea', a single full stop) are identified and counted as non-substantive.

**Normalisation.** The texts are converted to lower case, accents are removed via Unicode decomposition (unicodedata module) and apostrophes are replaced with spaces. This normalisation ensures that comparisons are not affected by the spelling variations common in responses typed quickly (missing accents, typographical or straight apostrophes).

**Characteristic terms.** TF-IDF weighting (Spärck Jones, 1972) highlights terms that are frequent within a response but rare across the entire corpus. It is calculated using scikit-learn on unigrams and bigrams of three or more letters, present in at least two responses, after removing a list of 115 French stopwords compiled for the study. The selected vocabulary comprises 163 terms. The weight of a term  $t$  in a response  $d$  is:

$$tfidf(t, d) = tf(t, d) \times \left[ \ln \left( \frac{1 + n}{1 + df(t)} \right) + 1 \right]$$

where  $tf(t, d)$  is the frequency of the term in the response,  $n$  is the number of responses and  $df(t)$  is the number of responses containing the term; each vector is then normalised. The terms are ranked according to their average weight across the corpus.

**Thematic coding.** The coding is deductive: eight themes were defined on the basis of the theoretical framework and an initial reading of the responses, then translated into explicit rules (regular expressions) applied automatically to the normalised text (Table 5). A single student may fall under several themes. The themes are grouped into three categories: task design,

control and supervision, and the conditions of the test or the framework for the AI.

**Table 5:** Thematic dictionary used for coding open-ended responses.

| Theme                                 | Category           | Examples of indicators sought                                                                                    |
|---------------------------------------|--------------------|------------------------------------------------------------------------------------------------------------------|
| Reasoning and justification           | Conceptualisation  | justification, reasoning, argument, reflection, critique, analysis, synthesis, taking a stance, explanation      |
| Contextualisation and personalisation | Conceptualisation  | personalised, experience, context, unique, situation, concrete, in class, during the lesson, authentic, original |
| Skills rather than rote learning      | Conceptualisation  | skills, by heart, memorisation, recall, abilities, technical skills, interpersonal skills                        |
| Projects, practice and process        | Conceptualisation  | project, practical work, practicals, practice, process, stages, monitoring, continuous, development, formative   |
| Diversification and hybridisation     | Conceptualisation  | diversify, combine, hybrid, mix, vary, portfolio, distribute                                                     |
| Oral examination and defence          | Examination format | oral, defence, interview, face-to-face, presentation, present, defence                                           |
| Supervised use of AI                  | AI framework       | integrating or supervising AI, authorising its use, AI cited as a tool, using AI, Copilot                        |
| Supervision and monitoring            | Control            | monitoring, in person, by telephone, via device, prohibiting use, sanctions, network disruption, on paper        |

*Note. The indicators are searched for in the standardised text (lower case, without accents).*

**Coding validation.** The dictionary was refined in two stages. The first stage revealed incorrect classifications, for example a response referring to ‘unauthorised uses of AI’ mistakenly classified under ‘supervised use’, or the sentence ‘the idea is not to ban AI’ classified under ‘monitoring’. The rules were

amended to exclude these cases. All final classifications were exported to a control file (codage\_themes.csv), which cross-references each response with its themes, and were manually reviewed in accordance with the verification protocol for thematic analysis (Braun and Clarke, 2006).

**Echo index.** To identify any responses that closely mirror the wording of the question, we calculate for each response to question 27 the proportion of the question’s eleven full words (e.g. ‘design’, ‘circumvent’, ‘fair’, ‘generative’) that it repeats. A value of 0.6 or higher indicates a high degree of repetition of the question’s wording. This index is an exploratory indicator and in no way constitutes evidence of AI use.

### 3.10. Software environment and reproducibility

The analyses were carried out in Python 3.13 using pandas 3.0, NumPy 2.5, SciPy 1.18, scikit-learn 1.9 and Matplotlib 3.11, in a Jupyter notebook that can be run from start to finish. All randomised steps use a fixed seed. The notebook, the anonymised data and the thematic coding file accompany the article: running the notebook on the source file reproduces every figure, table and chart exactly as shown.

### 3.11. Quality criteria for the case study

Three mechanisms underpin the rigour of the study. Firstly, methodological triangulation: a single outcome – scepticism towards AI – is examined using descriptive statistics, a paired t-test, reliability analysis, classification and analysis of open-ended responses. Secondly, traceability: every data processing decision (exclusions, recoding, parameters, coding rules) is recorded in the code and can be audited. Finally, transparency: the limitations of the study design, notably the sample size and the self-reported nature of the data, are discussed in section 6.1.

## 4. Results

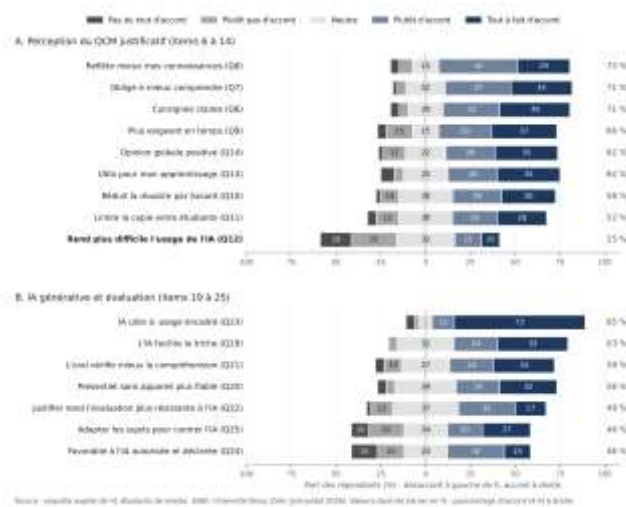
### 4.1. Widely recognised educational value

Students attribute clear educational value to the justifiable multiple-choice questionnaire. 73 per cent believe it better reflects their actual knowledge than a conventional multiple-choice questionnaire (item 8,  $M = 3.85$ ), and 71 per cent believe it forces them to understand the content rather than simply recognise the correct answer (item 7,  $M = 3.95$ ). 62 per cent have a generally positive opinion of them (item 14,  $M = 3.80$ ). When asked about the main advantage of the format, 30 out of 41 students (73 per cent) cited an educational benefit: the development of reasoning and critical thinking (17) or a

better assessment of understanding (13). Only 7 (17 per cent) cited the prevention of cheating or the use of AI.

The main cost is time. 66 per cent consider the exam to be more time-consuming (item 9), and 25 of the 39 students who answered question 16 cited this as the main drawback. The subjectivity of marking (4) and stress (3) remain marginal.

The nine perception items form a reliable scale ( $\alpha = 0.82$ ). One item, however, is an exception: item 12, ‘this format makes it more difficult to use AI’, shows a near-zero item-total correlation ( $r = 0.11$ ), and its removal raises the alpha to 0.86. Statistically, therefore, perceived resistance to AI does not fall within the same dimension as educational value. A seven-item educational value scale (items 6, 7, 8, 10, 11, 13, 14;  $\alpha = 0.85$ ) is retained for further analysis.



**Figure 2:** A format considered educational, but offering little protection against AI.

*Note.* Distribution of responses to the Likert items. 25 per cent of students believe that the format makes AI more difficult to use (item 12); 71 per cent feel that it requires a better understanding (item 7).

**4.2. The paradox: a format that fosters learning but offers little protection against AI**

Item 12 received the lowest score in the questionnaire ( $M = 2.75$ ). Only 25 per cent of students believe that the justifiable multiple-choice test makes AI more difficult to use, whilst 42.5 per cent believe the opposite. The discrepancy with the pedagogical value is significant amongst these same students, with strong effects.

**Table 6:** Comparison of educational value and perceived resistance to AI (Wilcoxon signed-rank tests).

| Paired comparison                                                                                      | Median values | <i>W</i> | <i>p</i> | <i>r</i> |
|--------------------------------------------------------------------------------------------------------|---------------|----------|----------|----------|
| Requires a better understanding (7) vs resists AI (12)                                                 | 4 vs 3        | 39.0     | < 0.001  | 0.62     |
| Reduces success by chance (10) vs resists AI (12)                                                      | 4 vs 3        | 51.5     | < 0.001  | 0.53     |
| Justification generally makes the assessment more robust (22) vs this format is robust against AI (12) | 3.5 vs 3      | 85.5     | 0.001    | 0.52     |
| Useful for my learning (13) vs. resistant to AI (12)                                                   | 4 vs 3        | 97.0     | 0.002    | 0.50     |
| Limits copying between students (11) vs resists AI (12)                                                | 4 vs 3        | 49.0     | 0.002    | 0.49     |

*Note:*  $n = 40$  pairs.  $r$  = effect size ( $|Z| / \sqrt{N}$ ).

Three findings shed light on this paradox. Firstly, the overall positive opinion (item 14) is strongly correlated with comprehension ( $\rho = 0.55$ ), fidelity to knowledge ( $\rho = 0.54$ ) and usefulness for learning ( $\rho = 0.50$ ;  $p \leq 0.001$ ), but not with resistance to AI ( $\rho = 0.21$ ;  $p = 0.18$ ). Secondly, students have greater confidence in the principle of justification (item 22: 49 per cent in agreement) than in this specific format (item 12: 25 per cent). Finally, of the 31 students who assign a high educational value to the format (score  $\geq 3.5$ ), only 8 believe it is resistant to AI.

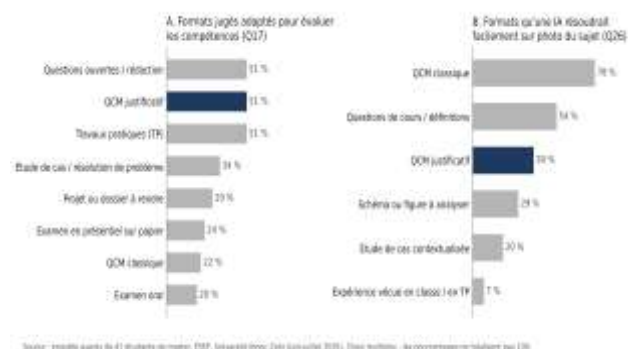
This scepticism does not depend on familiarity with AI: the reported frequency of use is not correlated with either item 12 ( $\rho = -0.05$ ) or the overall opinion ( $\rho = -0.13$ ).

**4.3. Assessment formats deemed suitable, and those deemed vulnerable**

The multiple-choice question with justification ranks among the three formats most often deemed suitable for genuinely assessing skills (51 per cent), on a par with open-ended questions and practical assignments. It is often associated with practical assignments: 13 students ticked both (Jaccard index = 0.45, the strongest association observed). At the same time, 39 per cent believe that an AI could easily solve a photographed multiple-choice question requiring justification.

Eight of the 21 students who consider it suitable also consider it vulnerable.

The hierarchy of perceived vulnerability is clear. It ranges from traditional multiple-choice questions (78 per cent) and questions on course material (54 per cent) to contextualised case studies (20 per cent) and, above all, questions linked to a real-life experience in class or during practical work (7 per cent). For students, therefore, what protects against AI is not the justification itself, but the fact that the question is rooted in a context with which the AI is unfamiliar.



**Figure 3:** The multiple-choice question requiring justification: one of the most popular formats, but considered vulnerable.

*Note.* Assessment formats considered suitable for assessing skills (A) and formats that an AI could easily solve using a photograph of the candidate (B). Multiple-choice.

More broadly, 40 out of 41 students believe that generative AI should lead to changes in assessment, with 17 stating it should lead to ‘profound’ changes. 63 per cent consider that it facilitates cheating (item 19). However, 85 per cent consider it useful if its use is regulated (item 23), but only 46 per cent would be in favour of an exam where it is permitted and declared (item 24). AI is accepted as a learning tool, but far less so as an assessment tool.

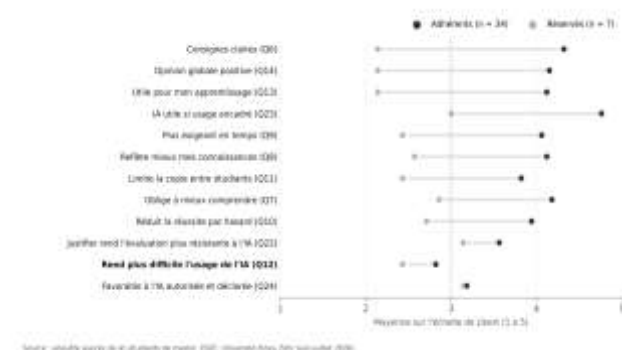
#### 4.4. Two student profiles

The k-means clustering reveals two profiles (silhouette = 0.33 for  $k = 2$ , compared with a maximum of 0.12 for  $k = 3$  to 5). The solution is robust: it agrees with Ward’s clustering (adjusted Rand index = 0.89) and remains stable across 200 initialisations (average ARI = 0.61).

- **The supporters** ( $n = 34$ , 83 per cent) consider the format to be clear, educational and fair (average educational value score = 4.08). They are also the most in favour of supervised use of AI (4.76).

- **Those with reservations** ( $n = 7$ , 17 per cent) reject the format on almost every front (score = 2.43) and remain neutral on AI (3.00).

The two groups differ by 1.5 to 2 points on most items, but not on resistance to AI: 2.82 amongst the supporters, 2.43 amongst the reserved group ( $p = 0.53$ ), both scores below the neutral point. Scepticism towards AI is therefore shared by the whole group, regardless of their assessment of the format. The profiles do not differ in terms of the frequency of AI use.



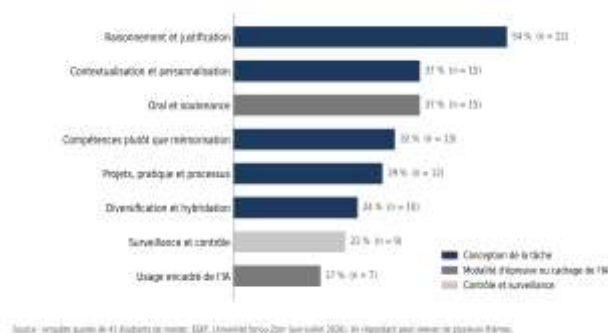
**Figure 4:** Two opposing profiles regarding educational value, both sceptical about AI.

*Note.* Item-by-item means for the two profiles derived from k-means clustering ( $k = 2$ , silhouette 0.33; agreement with Ward: ARI = 0.89).

#### 4.5. Students’ suggestions: designing rather than monitoring

Responses to question 27 had a median of 13 words (ranging from 2 to 87). 25 out of 41 students provided a substantive comment in response to question 28. The most characteristic terms according to TF-IDF weighting are ‘questions’, ‘exams’, ‘assessment’, ‘oral’, ‘justification’, ‘lessons’, ‘analysis’, ‘class’, ‘essay’, ‘projects’, ‘critique’ and ‘personal reflection’.

Thematic coding shows that 29 students (71 per cent) proposed at least one approach to task design, compared with 9 (22 per cent) who suggested an approach to assessment. Only 2 students focused solely on supervision. Reasoning and justification were the most common themes (54 per cent), followed by contextualisation and oral assessment (37 per cent each).



**Figure 5:** Students focus more on task design than on supervision.

*Note.* Automatic coding using a dictionary of responses to questions 27 and 28; a respondent may fall under several themes.

Some proposals are particularly elaborate. One student describes a ‘flash defence’: the student submits a written piece and then defends it in a 15-minute face-to-face session (E40). The same student proposes a two-part exam, first without AI and then with AI, the second set of questions containing a contradiction that only a student who has understood the material can spot. Others suggest a portfolio-based approach that assesses the process rather than the product (E09), materials provided only on the day of the exam (E25), or instructions asking students to compare their answer with that of an AI assistant (E31). In contrast, one student describes widespread cheating in the first semester and calls for stricter supervision (E17).

## 5. Discussion

### 5.1. Two functions to be distinguished: teaching and protection

The central finding of the study is a distinction. Students recognise that multiple-choice questions requiring justification have significant educational value, consistent with the literature on the justification of answers (Nield and Wintre, 1986; Dodd and Leal, 1988) and with the idea that assessment requiring reasoning promotes deep learning (Biggs, 1996; Scouller, 1998). However, they do not perceive it as a bulwark against AI. The reliability analysis confirms this: item 12 stands out statistically from the scale, and the classification shows that this scepticism cuts across both profiles.

This dissociation is consistent with what is known about current language models. A brief justification of a few lines is precisely the type of text that a model can easily produce, particularly based on a photographed subject (Susnjak and

McIntosh, 2024). Students, who are regular users of these tools, seem to have a practical awareness of this. The justification-based multiple-choice question must therefore be defended for what it does well — training and assessing reasoning — and not as a solution to the issue of integrity.

### 5.2. Contextual anchoring: a key factor in resistance

The hierarchy of perceived vulnerability (Figure 3) indicates where students perceive resilience to lie: in the task’s proximity to a context with which AI is unfamiliar. Only 7 per cent consider a question linked to a real-life experience in class or during practical sessions to be vulnerable, compared with 39 per cent for the justification-based multiple-choice question. Open-ended questions point in the same direction: contextualisation, personalisation, and documents provided on the day of the exam.

This finding aligns with the dimensions of authentic assessment described by Gulikers et al. (2004): the context of the task and the environment in which it is carried out. It suggests a simple approach for teachers: retain the multiple-choice with justification format, but ground the questions in what actually happened in class or during a practical session (an observation, a discussed error, an experimental result). It then becomes difficult to provide a justification without having actually experienced the session.

### 5.3. Designing rather than supervising

Seven out of ten students propose design approaches, whilst only two rely solely on supervision. These future teachers and trainers thus spontaneously adopt the approach to task redesign advocated by Lodge et al. (2023) and Perkins et al. (2024). Several of their proposals are directly applicable: a short oral defence of a written piece, a two-stage examination with and without AI, a process portfolio, and a critical comparison with a generated response.

The discrepancy between the acceptance of AI as a learning tool (85 per cent) and its low acceptance in examination settings (46 per cent) also merits attention. It undoubtedly reflects a concern for fairness: if AI is permitted, there is no guarantee that everyone will use it in the same way. An explicit framework setting out permitted levels of use, such as the AIAS scale, could address this concern.

## 5.4. Implications for exam design

**Table 7:** Design recommendations based on the findings.

| Principle                                     | Possible implementation in a multiple-choice question requiring justification                                             |
|-----------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| Ground the questions in the course experience | Questions relating to a specific session, practical or group observation                                                  |
| Assess the reasoning, not just the answer     | A marking scheme focused on justification, with explicit criteria                                                         |
| Combine assessment methods                    | A multiple-choice questionnaire requiring justification on paper, supplemented by a short oral response or a presentation |
| Reduce the time required                      | Short justifications (two or three lines) and an adjusted number of questions                                             |
| Setting clear guidelines for the use of AI    | Specify for each exam what is permitted, prohibited or must be declared                                                   |

## 5.5. The contribution of the computational approach

With a small corpus, the value of Python lies not in the volume processed but, in the rigour, and traceability. The reliability analysis revealed that item 12 measures something other than educational value, which a simple reading of the averages would not have shown. The classification made it possible to verify that scepticism towards AI is not limited to a single subgroup. Automated coding made the rules for assigning themes explicit and verifiable. Finally, the textual analysis identified two responses containing more than 60 per cent of the words from the question, suggesting that some open-ended responses may themselves have been drafted with the aid of AI. This point illustrates, within the data collection process itself, the phenomenon under study.

## 6. Limitations, future directions and conclusion

### 6.1. Limitations

The results apply to the case studied and do not claim to be statistically generalisable. The 82 per cent response rate limits non-response bias, but the sample remains small: the profile of the ‘reserved’ group ( $n = 7$ ) and comparisons between profiles are exploratory. The data are self-reported and relate

to perceptions, not to the actual performance of the AI in relation to the examination subject. As the questionnaire was distributed by the module tutor, a social desirability bias cannot be ruled out, although anonymity and the four-month interval following the examination limit its impact. Dictionary-based coding, which was manually checked, remains susceptible to unexpected phrasing. Finally, the possibility that some open-ended responses may have been drafted with the aid of AI cannot be ruled out.

### 6.2. Future directions

Three avenues for further research are envisaged. The first involves submitting a photograph of the actual exam paper to several multimodal models and marking their responses using the marking scheme applied to the students, in order to compare perceived vulnerability with measured vulnerability. The second involves designing and then testing a contextualised version of the multiple-choice question (MCQ) with justification, rooted in the module’s practical sessions. The third concerns a tool to assist in the generation of contextualised questions, within the field of educational technology.

### 6.3. Conclusion

This case study examined how Master’s students perceive the justification-based multiple-choice questionnaire at a time when generative AI is revolutionising assessment practices. It shows that these students clearly value this format for its impact on understanding and reasoning: seven out of ten believe it forces them to understand the content better, and the majority retain a positive opinion of it despite the time it requires. However, they do not regard it as offering any particular resistance to generative AI. Only a quarter of them believe it makes AI more difficult to use, and this scepticism is shared by the whole group, regardless of their views on the format or the frequency of AI use.

This finding suggests a distinction should be made between two functions that are often conflated: training and protection. The multiple-choice question requiring justification fulfils the former well, as it makes reasoning visible and limits success down to chance. It fulfils the latter poorly, as a justification of a few lines is precisely the type of text that a language model can produce with ease. From the students’ perspective, what provides protection is not the justification itself, but the fact that the question is rooted in a context unfamiliar to AI: what has been observed, discussed or experimented with in class and during practical sessions.

For exam designers, the lesson is twofold. Firstly, the multiple-choice question requiring justification should be retained for its formative value, streamlining it where necessary to reduce the time it takes to complete. Secondly, the focus should be on contextualising questions, diversifying assessment methods (written, oral, project-based) and explicitly setting out the framework for the use of AI, rather than relying solely on invigilation. The students' suggestions are also in line with this approach: seven out of ten favour redesigning tasks, whilst only a minority rely on monitoring.

From a methodological perspective, the study also demonstrates the value of an analysis powered by Python, even on a small corpus. Reliability analysis, classification and the automated coding of open-ended responses made it possible to go beyond simply examining averages, to verify the robustness of the results and to ensure that each step was traceable and reproducible. This approach can be applied to other modules and other institutions.

These conclusions remain specific to the case studied and are based on self-reported perceptions. They call for further experimental research, which would compare the perceived vulnerability of the assessment formats with their actual vulnerability when faced with current models. In the longer term, the challenge goes beyond the issue of cheating: it is a matter of designing assessments that measure what students have actually understood, in an environment where AI is now a standard part of learning tools.

## References

- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364. <https://doi.org/10.1007/BF00138871>
- Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment & Evaluation in Higher Education*, 30(1), 65–72. <https://doi.org/10.1080/0260293042003243904>
- Cotton, D. R. E., Cotton, P. A., and Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Dodd, D. K., and Leal, L. (1988). Answer justification: Removing the ‘trick’ from multiple-choice questions. *Teaching of Psychology*, 15(1), 37–38. [https://doi.org/10.1207/s15328023top1501\\_8](https://doi.org/10.1207/s15328023top1501_8)
- Gulikers, J. T. M., Bastiaens, T. J., and Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, 52(3), 67–86. <https://doi.org/10.1007/BF02504676>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Lodge, J. M., Howard, S., Bearman, M., Dawson, P., et al. (2023). Assessment reform for the age of artificial intelligence. *Tertiary Education Quality and Standards Agency*.
- McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt and J. Millman (eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- Moorhouse, B. L., Yeo, M. A., and Wan, Y. (2023). Generative AI tools and assessment: Guidelines from the world's top-ranking universities. *Computers and Education Open*, 5, Article 100151. <https://doi.org/10.1016/j.caeo.2023.100151>
- Nield, A. F., and Wintre, M. G. (1986). Multiple-choice questions with an option to comment: Student attitudes and use. *Teaching of Psychology*, 13(4), 196–199. [https://doi.org/10.1207/s15328023top1304\\_6](https://doi.org/10.1207/s15328023top1304_6)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

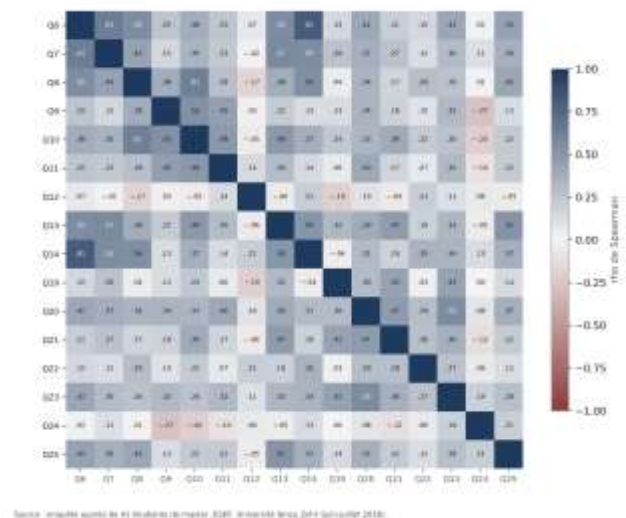
A., Courneau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

- Perkins, M., Furze, L., Roe, J., and MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for the ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, 21(6). <https://doi.org/10.53761/q3azde36>
- Romero, C., and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601–618.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Rudolph, J., Tan, S., and Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1), 342–363. <https://doi.org/10.37074/jalt.2023.6.1.9>
- Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple-choice question examination versus essay assignment. *Higher Education*, 35(4), 453–472. <https://doi.org/10.1023/A:1003196224280>
- Siemens, G., and Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). ACM. <https://doi.org/10.1145/2330601.2330661>
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>
- Susnjak, T., and McIntosh, T. R. (2024). ChatGPT: The end of online exam integrity? *Education Sciences*, 14(6), Article 656. <https://doi.org/10.3390/educsci14060656>

- Ward, J. H. (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association*, 58(301), 236–244. <https://doi.org/10.1080/01621459.1963.10500845>
- Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment, Research, and Evaluation*, 2, Article 2.
- Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.

## Appendix. Availability of data and code

The Jupyter notebook `analyse_qcm_justificatif.ipynb`, the anonymised data (`donnees_master_anonymisees.csv`) and the thematic coding file (`codage_themes.csv`) enable all the results to be reproduced. The complete correlation matrix is presented below.



**Figure A1:** Spearman's correlation matrix between Likert items.

*Note.* Spearman's rho coefficients between the sixteen Likert items (n = 40 to 41).