Vol.8 No.1 (2025)

Journal of Applied Learning & Teaching

ISSN : 2591-801X

Content Available at : http://journals.sfu.ca/jalt/index.php/jalt/index

Appraising higher education assessment validity: Development of the PANDORA GenAI Susceptibility Rubric

Peter Bannister[A] A *Researcher, Universidad Internacional de La Rioja, Spain*

Alexandra Santamaría Urbieta[B] B *Professor, Universidad Internacional de La Rioja, Spain*

Nuria Brufau Alvira[C] C *Lecturer, Universidad Internacional de La Rioja, Spain*

Keywords

Academic integrity;
assessment design;
generative ai;
higher education;
susceptibility rubric.

Correspondence

peter.bannister@unir.net [A]

Article Info

Received 16 October 2024
Received in revised form 15 January 2025
Accepted 17 January 2025
Available online 23 January 2025

DOI: https://doi.org/10.37074/jalt.2025.8.1.20

PDF Link to **Rubrics**

Abstract

This paper presents the development and application of the PANDORA GenAI Susceptibility Rubric, a novel tool designed to assess the susceptibility of higher education assessments to the undeclared use of generative artificial intelligence (GenAI) tools. In response to growing concerns about academic integrity and the rising sophistication of GenAI technologies, the rubric provides educators with a structured framework to critically evaluate the validity of their assessments across key criteria, including collaborative authorship, intellectual task complexity, and the opportunity for creativity. Through a mixed-methods design, the rubric was refined to include expert-informed modifications and validated through end-user application across various arts and humanities courses. Results highlight how assessment design can either mitigate or exacerbate GenAI susceptibility, revealing that tasks requiring genuine collaboration, creative thinking, and process-oriented evaluation offer greater resistance to AI manipulation. The rubric also emphasises balancing detailed guidance with student autonomy to avoid facilitating GenAI prompt formulation. This study contributes to the field by offering a practical instrument that promotes more robust, ethically sound, and future-proofed assessment practices. It serves as a critical response to the pressing challenges posed by GenAI in higher education and informs ongoing discourse on academic integrity.

Introduction

The discourse surrounding the disruptive impacts of human interaction with progressively capable Generative Artificial Intelligence (GenAI) technologies in higher education (HE) has broadly centred around the challenges these pose to academic integrity and, to a lesser extent, affordances for teaching and learning (Bond et al.; 2024; Bozkurt; 2023; Law, 2024). Initial muted framings of the duality of GenAI as a "bullshit spewer" and harbinger of traditional assessment approach demise (Rudolph et al., 2023, p. 342) have gained traction. Nonetheless, other analogous comparisons to former technological innovations, such as the calculator, seem to have come up short regarding the overarching complexity which these technologies introduce (Lodge et al., 2023) in the little over two years since the release of OpenAI's ChatGPT 3.5 in November 2022. Their subsequent, more capable iterations have produced additional waves of impact. These have cemented GenAI globally as a highly complex matter which continues to defy initial reductionist underestimations of the technology's potential (Möck, 2022).

In parallel, scholars have come together to document differing perspectives on the perils and promises. A swift Google Scholar search reveals almost 6000 publications address GenAI challenges and opportunities, between 2022 and the time of writing. Whilst the remit of the present paper resides in a more specific focus, it is worthy to note that much of this work still refers to this particular phenomenon as 'the impact of GenAI in HE' in the singular, as if each discipline of the academy were a singular, homogenous, monolithic entity. Here, the authors prefer to refer to 'the *impacts* of GenAI' in *plural*. This is so, not solely as a testament to the rich diverse plurality of the sector itself but also considering that each novel germination of GenAI tools fundamentally changes the rules of the game and will continue to do so time and time again going forward (Bannister & Carver., 2024). The very nature of this multifaceted disruption to HE is pervasive not only to shining a light on longstanding inefficiencies in established assessment practice but also invites us to reconsider more profound epistemological, ontological, and axiological assumptions, and ultimately, the very purpose of HE itself (Kramm & McKenna, 2023).

As these complexities emerge, it becomes clear, however, that abstract considerations cannot overshadow the immediate practical challenges faced by educators on the ground, for instance, the practicalities of adapting assessment design. Further work needs to be undertaken to collate and critically examine domain-specific examples of good practice (e.g., Bannister et al., 2023). These ought to be robust enough to traverse the tensions of the quest for authenticity in assessments whilst deliberating on the "hauntological turn", referring to the ways in which AI can act as "ghostly forces shaping knowledge in unseen ways, paradoxically omnipresent and absent" (Vallis, 2024, p. 4-5), in the operationalisation of synthetic simulations of student output produced by GenAI tools. Considering these multifaceted challenges, there is a pressing need for structured frameworks to guide educators through this complex terrain.

Perkins et al.'s (2024) Artificial Intelligence Assessment Scale (AIAS) offers an excellent starting point to this end which has been positively received in differing spheres of education globally. In contrast to many techno-teratological punitive approaches found elsewhere, in their constructive and ethical reframing of GenAI tool usage, , the authors strive to strike a balance between pedagogical benefits and ethical concerns. The AIAS serves as a structured mechanism which supports educators in this novel and, at times, pernicious terrain that advocated for greater transparency and responsibility through the articulation of clear guidance for all key stakeholders, often found to be lacking in the assessment process (Bearman & Luckin, 2020; Nguyen et al., 2024). The gradual incrementation across the instrument is a poignant reminder that while GenAI can enhance certain aspects of assessment, its inclusion is not necessarily inevitable or compulsory. Ultimately, educators are encouragingly framed as agentive architects in the assessment design process, as opposed to reactive subjects of impervious technological change.

Further complexity can be found in praxis. Mounting empirical validation gives credence to nascent media speculation with even ChatGPT 3.5 having been found to have produced satisfactory performance in varying assessments across subject domains (Lo, 2023). More recent research consolidates this with a confirmation of a continued positive incremental trend following more potent iterations of these technologies which achieve parity with their human-equivalent test-takers across a broad spectrum of disciplines at differing educational levels (e.g., Newton & Xiromeriti, 2024). As GenAI demonstrates increasing proficiency in producing passable academic assessment outputs (Smolansky et al., 2023), the question of how to detect its undeclared use adds another layer of complexity to the assessment landscape. Be that as it may, calls for substantive assessment design reform and innovation are longstanding fixtures of extant literature, albeit thus far with highly limited permeation in praxis and policy (Boud & Falchikov, 2007; Morris et al., 2021; Sadler et al., 2022, amongst many others), even after the upheavals of the disruptive COVID-19 pandemic (Kaqinari, 2023). The essay, for instance, is the mainstay of many domains (Meylani, 2024). It is the traditional assessment genre par excellence, particularly in disciplines such as the Arts and Humanities (Zemits, 2017), the context in which this study is grounded. For some, an essay can be pedagogically beneficial as a focused act of intellectual exploration in which theoretical matters addressed in the learning process may be practically applied through the prism of critical thinking (West et al., 2019). Scholarly consensus, however, converges with dissentious standpoints which criticise surface learning strategies and formulaic responses employed (McGraw & Mason, 2021). Further criticisms encompass the passive reproduction of the ideas of others with no authentic critical discernment (Mirador, 2018), the potential for a lack of alignment with intended learning outcomes, future professional realities (Lalor et al., 2015), extraneous cognitive load and assessment anxiety (Teixeira et al., 2022), grading subjectivity (Bloxham et al., 2011), inconsistency (Lipnevich et al., 2020) and, bias and cultural disparities amongst increasingly international student cohorts (Melekhina & Levitan, 2015).

This genre is particularly susceptible to the undeclared use of GenAI tools for unfair advantage. This is so, as the tools are adept at generating content, written at a standard higher than that of some human learner-produced work (Herbold et al., 2023) in the form of coherent long-form text that fulfils surface-level requirements of often vague essay questions (Waltzer et al., 2024), notwithstanding exceptions of hallucination (Alkaissi & McFarlane, 2023). Egloff (2024) frames this as a tempting shortcut as opposed to the more cognitively demanding task of meaningfully engaging with the material. Furthermore, as Sharples (2022) remarks, if these technologies are mindlessly applied for the creation of the entire piece and lecturers subsequently use the same or similar tools for review, "nobody learns, nobody gains" (p. 1124). This admittedly dystopian vision would render the HE experiences as little more than a mechanically transactional exercise of credentialling devoid of meaningful intellectual growth. As such, students-as-consumers go through the motions and yet do not experience the personally transformative potential of a tertiary education first-hand (Ashwin, 2020). Ashwin et al. (2023), in their timely return to this matter, advocate for a vision of HE in which students engage in a dynamic relationship with knowledge framed as playing a crucial role in their future career prospects. On balance, owing to the imminent need to furnish students with future skills and graduate attributes for a professional world in which GenAI will progressively become more enmeshed (Ehlers & Eigbrecht, 2024; Pratschke, 2024), attempts to eradicate or implement punitive detection in assessment (Blackie, 2024) in the interim are a counterintuitive act of folly.

The thorny issue of detection in practice thus far, in both its human and technologically led facets, stands out as both ineffective and inadequate. Perhaps owing to the continued absence of an effective established protocol for educator AI-generated text decipherment (Kirmani, 2023), multiple studies attest to rates of limited educator success in weeding out authentic assessment artefacts of sole human authorship (Fleckenstein et al., 2024; Matthews & Volpe, 2023; Liu et al., 2023; amongst others). Despite cumulative scholarly evidence to the contrary, commercial enterprises persist in their claims regarding AI-generated text detection efficacy with limited scholarship diverging (Walters, 2023) from the consensus to the contrary (Baron, 2024; Chaka, 2024; Kumar & Mindzak, 2024; Weber-Wulff et al., 2023). Perkins et al. (2024) illustrate a compendium of adversarial techniques which users can use to manipulate GenAI outputs tensuring the probability of going under the radar of text classifiers is increased. The sensitivities of authorial linguistic traits under the text classifier microscope were also examined by Liang et al. (2023) who notably sounded the alarm regarding the inequitable susceptibility to false positives of work produced by authors who use English as an additional language. Deepening concerns here, existing GenAI academic integrity policies often fail to address the specific needs of this demographic, despite their substantial academic, cultural, and lucrative financial contributions in neoliberal HE models (Bannister et al., 2024b). These highly problematic issues in Higher Education Institution (HEI) response are illustrative of additional complexity at play, which thus far in practice seems to have largely failed to consider the wide diversity of students on campus and their bespoke characteristics and needs, not all of which can be blamed on the technology itself. Rather, these matters expose the inadequacies of institutional responses that overlook the intricate interplay of human, cultural, and systemic factors in the engagement with GenAI.

Beyond the technical challenges of detection lies a more insidious issue: the uncritical acceptance and propagation of certain narratives surrounding GenAI in education. In line with Monett and Grigorescu (2024), the 'GenAI in education mythogenesis' in popular discourse and in scholarship speaks to an undiscerning acquiescence that ought to be challenged. For instance, the personalisation of learning and democratisation of access to knowledge are AI-generated tropes which, with their notably American English spelling, often adorn GenAI written academic discourse with little scrutiny, albeit with some exceptions (e.g., Tafazoli, 2024). And yet, they inadvertently mask the deeper inequalities they perpetuate, raising questions about whose knowledge is being privileged, whose learning paths are being pre-determined, and the new digital divide with the most potent technologies only accessible in exchange for payment which not all can afford. Further concern in scholarship also pertains to eyebrow-raising instances of publications which contain evidently apparent passages of text of GenAI chatbot responses as highlighted by Giray (2024). The issue here is not at all to do with the use of GenAI tools as a dialogic idea generator, text perfector, or even research companion and/or co-constructor (Tang et al., 2024). It is rather the unsettling dereliction of duty and undermining lack of rigour not only of the authors but of all those involved which make a mockery of the publication process. This sheds light on longstanding systemic issues that still need to be addressed in the publication process even today, in the so-called age of AI. These musings aim to broaden the scope of GenAI-empowered academic misconduct discourse to encompass a more comprehensive understanding of this complex phenomenon, which serves as a reminder to avoid the reductive fallacy of attributing the entirety of the issue solely to student behaviour.

In this spirit, amidst this maelstrom of challenges, in their seminal conceptual work, Dawson et al. (2024) address these emerging complexities through the proposal of a transformative paradigm shift in our approach to academic integrity concerns, which directly resonates with the broader issues articulated thus far. Rather than maintaining a narrow focus on cheating as a moralistic deviation, the authors argue for reframing the issue through the lens of assessment validity. They posit that the latter offers a more constructive and inclusive framework for understanding the educational consequences of GenAI usage. In linking validity to the fundamental purpose of HE, this reframing also aligns with the broader proposition that GenAI's impact on HE is not a monolithic phenomenon, but one that evolves with each technological iteration, thereby continually reshaping the parameters of academic integrity and assessment practices. Thus, the authors' call for a shift from moralisation to validity invites a critical reconsideration of both assessment design and implementation, and the very assumptions upon which our educational frameworks rest.

This bold proposition is a formidable undertaking, and it is precisely this call to action that has catalysed the development of this paper. It has come to fruition within the framework of *The PANDORA Project* spearheaded by the authors together with a team of scholars across Spain and the UK from 2023-2025 funded by Universidad Internacional de La Rioja (UNIR), Spain. This grassroots project has been conceptualised to analyse current assessment validity through the prism of AI-enabled academic misconduct in the Faculty of Arts and Humanities. Furthermore, it seeks to subsequently forge novel assessment innovations with and without AI integration and inform institutional guidelines which regulate the use of GenAI tools. While the authors do not purport to provide exhaustive solutions to either this or all the concerns detailed thus far, an initial step is taken here in this paper. Stemming from the first stage of the project, this paper documents the development of a novel evaluative tool for HEIs, grounded in the context of arts and humanities tertiary education, that seeks to measure the validity of assessments in the context of increasingly sophisticated GenAI tools.

Literature

Ambiguity and institutional inertia in a global regulatory vacuum

The HEI response to the complexities outlined has yielded limited and problematic guidance marked by critical silences (Luo, 2024), which fail to recognise the full repertoire of complexities of human-AI interaction in HE, as detailed previously, with many institutions internationally still struggling to establish definitive regulations (Dai et al., 2024; De Maio, 2024; McDonald et al., 2024; Sok & Heng, 2024). This enduring global regulatory vacuum, a concurrent sociotechnological phenomenon remains a point of contention, despite scholarly green shoots aimed at shaping effective and socially just policy development (e.g., Bannister et al., 2024a). Even UNESCO (2023), long regarded as a beacon for global educational policy and thought leadership, has been found to fall short with their *Guide to Generative AI in Education and Research* penned by Maio and Holmes in 2023. Whilst initially well-received, it has subsequently been criticised for vague recommendations that overlook the complex realities of diverse educational contexts and not providing detailed, actionable frameworks for implementation in different regions (Knight et al., 2023; Taylor, 2024). In this light, increased levels of educator technostress (Kohnke et al., 2024) are evidently comprehensible, although talk of a GenAI-fuelled crisis in HE (Song, 2024) might seemingly be dismissed as little more than a hyperbolic hissy fit prima facie (Leaver & Srdarov, 2023). However, on closer inspection, regulatory ambiguity and institutional inertia combined with GenAI-induced assessment validity erosion may indeed be indictive of a perfect storm in which the sector finds itself embroiled.

HE quality assurance and GenAI

In this landscape, quality assurance reiterates itself as a fundamental component of HE. Back in her now somewhat dated literature review, Ryan (2015) highlighted the importance of encouraging the implementation of global initiatives in the future trends section. Now over a decade later, on the issue of GenAI, these have been deemed to be highly necessary (Rawas, 2023) but are often found somewhat lacking by educators (Bannister, 2024). Furthermore, there are two key ideas of relevance to the genesis of the present study. On the one hand, that quality as a concept may be framed in differing ways each with multifaceted interpretations and, on the other, one such framing may be the pursuit of quality as an agent for transformative continuous improvement. Although the notion of accountability, i.e. the goal of quality assurance, and continuous improvement, i.e. the aim of quality enhancement, are often conceived as two incongruous forces, they can also be viewed as complementary that together create a balanced approach to educational development (Condette, 2024). Moreover, as Asiyai (2020) highlights, innovation plays a pivotal role to this end, and perhaps now more than ever if the assessment validity challenges are to be addressed and may act as a catalyst which unlocks this aforementioned transformative power. The operationalisation of this premise can be achieved through the forward-thinking creation of dynamic new tools and frameworks which are both able to assess the status quo and inform future directions (Kaiser et al., 2022). Despite ongoing efforts, existing frameworks lack the structural depth necessary to address the associated complexities. For instance, Zaphir et al.'s (2024) MAGE Framework addresses GenAI vulnerability of assessments. However, the study's focus primarily resides on the correlation between the quality of GenAI responses and engineered prompts and does not fully address academic misconduct concerns or assessment redesign approaches based on comprehensive vulnerability criteria. That's where this study endeavours to offer a more robust, actionable instrument for assessment GenAI susceptibility diagnosis as a basis to inform educators on GenAI-aware assessment design going forward.

Conceptualising assessment validity and GenAI susceptibility
Returning to assessment validity, this is a cornerstone concept in educational measurement which extends far beyond the rudimentary notion of an assessment measuring what it purports to measure (Dawson et al., 2024). In the context of HE, it encompasses the intricate interplay between the assessment design, its implementation, and its ability to authentically capture and evaluate the intended learning outcomes (Ajjawi et al., 2019). As articulated by Chapelle and Lee (2021), validity is not merely a property of the assessment itself, but rather an evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores. It is crucial to also acknowledge that the challenges posed by GenAI to assessment validity are not unprecedented but rather represent an intensification of longstanding concerns (Farrelly & Baker, 2023). Previous threats to assessment validity, while perhaps less technologically sophisticated, have been and continue to be equally pressing. For instance,

plagiarism, in its many forms (Eaton, 2017), is an issue which persists with unwavering resilience despite diverse comprehensive efforts globally to tackle this (Tight, 2024). Moreover, the proliferation of contract cheating services and essay mills has long cast doubt on the integrity of traditional assessment methods (Ahsan et al., 2021), with quite frankly tragic consequences for those who fall victim (Pitt et al., 2020).

These issues, while mediatised to a certain extent, failed to catalyse comprehensive reform in HE (Sweeny, 2023). However, it is only now, in the face of GenAI's unprecedented capabilities, that we observe a growing consensus across academic, administrative, and even corporate sectors on the urgent need to address these long-standing issues comprehensively (Perkins et al., 2024). This shift in perspective underscores the transformative potential of the current moment, where the challenges posed by GenAI may serve as the tipping point for a fundamental reassessment of how we conceptualise and implement valid assessments in HE (Dawson et al., 2024). Moreover, given the extensive discourse on GenAI-assisted academic misconduct, it is imperative to provide a more precise articulation of its implications for assessment validity, addressing both apparent and underlying issues. Building on the literature discussed, we propose a novel conceptual definition of 'GenAI susceptibility' that moves beyond moralistic notions of misconduct:

The undeclared use of GenAI to spectrally engineer a quasi- or fully synthetic AI-human hybrid output by any member of the academic community contrived to catalyse an erosion of ontological authenticity for epistemic refraction. In assessment validity, this unfair advantage renders such a submission a distorted reflection of knowledge acquisition, notwithstanding pre-existing assessment design flaws, which may inadvertently achieve comparable misrepresentations, ultimately corrupting the very raison d'être of the assessment itself. However, this issue transcends the confines of student-directed educational assessments, as the implications extend to broader academic spheres such as scientific writing authorship and scholarly publication, peer review, and intellectual integrity more broadly.

This is a working definition in need of further refinement. We encourage scholars to critically examine and expand upon this definition as GenAI technologies continue to evolve and HEIs adjust their responses accordingly.

Operationalising GenAI susceptibility for assessment validity
Given these concerns about validity and academic integrity, rubrics offer a potential solution for mitigating GenAI susceptibility in assessments. These instruments are traditionally framed within education as student-facing instruments that enhance the assessment process's transparency (Panadero et al., 2023). However, their potential extends beyond mere criteria clarification; they serve as mechanisms to enhance accuracy and reduce cognitive load (Krebs et al., 2022). They also foster self-regulated learning (Fraile et al., 2023) and critical engagement with evaluative standards (Francis, 2018), thus giving rise to the creation of pedagogical bridges, which can align learning objectives with assessment criteria and instructional strategies (Ragupathi & Lee, 2020). A further affordance resides in the linguistic commonality which can scaffold discussions around academic expectations and performance enhancements between stakeholders (Morton et al., 2021).

Nevertheless, there are documented associated risks in their use, such as the potential reductionist constraint of the creativity and higher-order thinking they aim to assess, particularly when applied mechanistically (Panadero & Jonsson, 2020). The prescriptive nature of rubrics in assessment may also reinforce power imbalances within educational settings according to Tan (2020). According to the author, this is so given that as artefacts of educational measurement, they embody particular ontological and axiological assumptions about what constitutes valuable learning and how it should be evidenced. This delimitation can risk privileging certain forms of knowledge expression over others and potentially marginalising diverse ways of knowing and demonstrating understanding.

Conversely, in the context of the PANDORA Project and the broader challenges posed by GenAI, the rubric developed here is conceived to assume a more nuanced and multifaceted role. The application of the traditional diagnostic function in the present context is repurposed to evaluate not the submitted assessment artefacts by test-takers, but rather to appraise the assessment design itself and its susceptibility to undeclared GenAI tool usage. In accordance with Bearman and Ajjawi (2019), the instrument is intended to embody a non-static role of enactment in which it is also conceived as a dialogical conduit which invites focused critical reflection which encourages consideration of aspects such as assessment alignment with intended learning outcomes, technological susceptibility, and the authenticity of tasks in relation to higher-order thinking skills. Ultimately, the impetus of the tool is to enable an informed decision to be taken by educators based on a structured analysis as to whether assessments truly capture the essence of student learning or inadvertently create loopholes susceptible to undeclared GenAI tool usage. This metamorphosis is intended to endow the rubric, whose development is reported here, with the capacity to serve as a reflective tool that acts as a dialogic catalyst for pedagogical introspection and innovation.

Although much of the literature on rubric construction is grounded in student-facing contexts, there are, however, a series of constructive insights in such studies as to key design principles and pitfalls to avoid, which are of relevance here. For instance, efficacious rubrics frequently incorporate unambiguous and specific criteria that align closely with the construct under evaluation (Boyd & Hill, 2024), whilst simultaneously providing descriptive levels that facilitate nuanced evaluation (Brookhart, 2018). Moreover, well-constructed rubrics engender consistency in evaluation (Bennett, 2016) and can be used as an impetus for substantive feedback to guide iterative improvement (Wilkerson, 2020). Conversely, prevalent pitfalls in rubric design encompass excessively vague or subjective language, potentially leading to inconsistent interpretations among evaluators (Kenworthy & Hrivnak, 2014). Furthermore, rubrics characterised by excessive complexity or granularity may prove unwieldy in practice, potentially diminishing their

pragmatic utility in the evaluation process (Perris & Mohee, 2020).

Drawing on this, the following section details the specific methodological approach taken for the development of the novel PANDORA GenAI Susceptibility Rubric.

Rubric development process

The mixed-methods developmental design architecture for the rubric took place across three interconnected collaborative phases: preliminary exploratory inductive analysis, expert review and iterative refinement, and end-user validation (Allen & Knight, 2009). In the interests of concision, information pertaining to the practical application of this procedure is detailed alongside the methodological particulars to dedicate the preponderance of the final sections of the paper to the resulting instrument and its potential implications for theory and practice.

Context and participants

This study was conducted internally at Universidad Internacional de La Rioja, Spain. It was specifically grounded in the context of Arts and Humanities, owing to the prevalence of traditional assessment methods in this discipline, such as the essay. It has been conceptualised as a pilot project with potential replicable scalability to the wider institution and cognate settings nationally and internationally. After launching a faculty-wide call for participants with the support of academic senior managers, 12 lecturers who taught one or more modules pertaining to the programmes of study highlighted below, were recruited:

- BA Translation and Interpretation Studies
- BA Art History
- MA Theatre Studies
- MA Cultural Management
- MA Teaching English as a Second Language
- MA Secondary Education Teacher Training

This range of programmes provided a representation of Arts and Humanities disciplines, which allowed for a nuanced exploration of GenAI susceptibility across different subject areas within the broader field. The inclusion of both undergraduate and postgraduate programmes enabled the researchers to examine potential variations in assessment practices and GenAI susceptibility at different levels of HE. The participation of educators from these varied programmes ensured a rich dataset, encompassing a wide array of assessment types, subject-specific considerations, and pedagogical approaches. This diversity was crucial in developing a robust and versatile rubric capable of addressing the multifaceted nature of GenAI susceptibility across the Arts and Humanities.

In the first instance, voluntary response sampling was used, however, owing to the highly limited responses to the open call sent by email to all Arts and Humanities faculty members, we turned to the snowball sampling technique (Cash et al., 2022) by asking initial participants to recommend further potential participants who were contacted directly. While not statistically representative of the entire faculty, the resulting sample provided a diverse range of perspectives from experienced educators across various arts and humanities disciplines. Participants were qualified to PhD level and had at least three years' university teaching experience. Most of the respondents were of Spanish nationality with two participants hailing from Venezuela and the United Kingdom. There was a female-to-male gender balance of ten to two.

Phase 1- Preliminary exploratory inductive analysis

Despite an unexpectedly somewhat muted response, which itself became a point of analysis potentially indicating a lack of awareness or engagement with AI-related issues in assessment practices, the research team proceeded with an exploratory inductive analysis. This entailed two stages which comprised semi-structured interviews with participants and qualitative content analysis of assessment task briefs, rubrics and artefacts of the modules they teach.

The semi-structured interviews were devised to serve a dual purpose: to gather preliminary reflections on the types of assessments employed in arts and humanities courses and to gauge educators' initial perceptions of their assessments' potential susceptibility to GenAI-assisted completion. A bespoke standardised interview protocol was developed by the research team, which comprised open-ended questions to ensure consistency across interviews while allowing for flexible exploration of emerging themes. The protocol included the following key areas of inquiry: current assessment practices, perceived challenges in design, and perceptions and experiences of GenAI's impact on HE assessments. Data was then coded systematically using thematic analysis (Braun et al., 2025) to identify recurring patterns and prevalent themes, such as 'confidence in assessment validity' and 'limited practical engagement with GenAI tools', resulting in 'limited informed awareness, related to the interaction between GenAI tool potentialities and academic assessment practices', as is illustrated from the translated participant quotes below:

> I'm confident that my assessments are still valid. I think I could spot anything that wasn't written by a human being anyway. After all, they require critical thinking and analysis, which AI can't replicate... can it? [CB57.09.G]

> To be honest, I haven't really engaged with these AI tools myself. I've heard a lot about ChatGPT, but I'm not sure how they could impact my assessments. [TH08.17.B]

> From what I've heard they haven't updated it with more recent information so my assessments on the new education law are safe. I can come back to this

again in future. [HB59.21.T]

Following the interviews, participating educators were invited to submit a range of materials, including assessment instructions, marking rubrics (where available), and anonymised samples of graded student work. The full repertoire of the 34 assessments collected is detailed in Table 1 below:

Table 1. Compendium of assessments analysed.

Level of Study	Degree Programme	Module	Language	Assessment Type
Undergraduate	BA Translation and Interpretation	Introduction to Translation	Spanish	Critical commentary
		History, Politics and Culture in Anglophone Countries	English	Essay
	BA Art History	Archaeology	Spanish	Essay
				Harris matrix
				Written justified identification of archaeological remains
		Baroque Art	Spanish	Post-museum visit critical commentary
				Forum post
				Baroque painting analytical written commentary
		Art from Neoclassicism to Avant-Garde	Spanish	Neoclassical painting analytical written commentary
				Group oral presentation
				Avant-garde painting analytical commentary oral recording
		Early Medieval and Romantic Art	Spanish	Painting analytical commentary oral recording
				Written historical contextualisation of works of art
				Forum post
		Modern Age Art in Spain	Spanish	Painting analytical commentary
				Group scientific article text commentary
				Essay
		Industrial Arts	Spanish	Essay
				Forum post
				Painting analytical commentary
Postgraduate	MA Cultural Management	Publishing Management	Spanish	Group publish project
				Book launch proposal
				Communication strategy dossier
		Current Challenges in Culture Management	Spanish	Fundraising strategic plan
				Institutional analysis
				Public institution and foundation comparative analysis
	MA Theatre Studies	Theatrical Pedagogy	Spanish	Forum post
				Research proposal
				Essay
	MA Teaching English as a Second Language	Language Systems	English	Lesson Plan
				Tutorial Roleplay Video
				Lesson Plan
	MA Secondary Education Teacher Training	Curricular Design in Physics and Chemistry	Spanish	Lesson plan
		Disciplinary Training in Physics and Chemistry	Spanish	Infographic shared on social networks

These materials formed the corpus for subsequent qualitative content analysis. Two research team members conducted this analytical process through joint coding procedures and subsequent discussions to resolve discrepancies.

To further evaluate the susceptibility of these assessments to undeclared GenAI-assisted completion, the two researchers input the assessment task briefs into two prominent GenAI tools, Open AI's ChatGPT-3.5 and Anthropic's Claude 3 Haiku. The aim was to determine whether responses generated would achieve a passable grade according to the provided assessment rubrics, to determine the degree of GenAI susceptibility of each assessment task. As a means of quantifying this, a 5-point Likert-type scale (1-5) was implemented, chosen for its balance of reliability, validity, and practical utility in educational technology assessment (DeVellis & Thorpe, 2021). Table 2 below illustrates how the construct was operationalised:

The third member of the research team subsequently carried out the same analyses independently, and outcomes from both were ratified with the following outcome as per Table 3 below:

Table 2. Initial GenAI susceptibility scale for assessment tasks.

Scale Point	Susceptibility Level	Description
1	Very Low	AI-generated response fails to meet basic task requirements
2	Low	AI-generated response partially fails to meet requirements but would fall short of a passing grade
3	Moderate	AI-generated response meets minimum requirements for a pass
4	High	AI-generated response would achieve a good grade
5	High	AI-generated response would achieve an excellent grade

Table 3. Preliminary assessment validity GenAI susceptibility findings.

Scale Point	Number of Assessments	% of Corpus
1. Very Low Susceptibility	0	0%
2. Low Susceptibility	1	2.94%
3. Moderate Susceptibility	5	14.71%
4. High Susceptibility	5	14.71%
5. Very High Susceptibility	23	67.65%

The complementary qualitative findings revealed that assignments with lower GenAI susceptibility shared several key attributes. Predominantly, these tasks required direct student engagement or physical presence, such as museum visits or video creation, and often involved practical applications or hands-on tasks rather than purely theoretical or text-based work. For instance, tasks rated at GenAI Susceptibility Level 2 (low susceptibility) required the creation of videos, a format that presents significant challenges for GenAI to replicate at the time of writing. Similarly, these tasks involved commenting on specific, non-named images, reducing the ease with which AI could complete the assignment.

In assignments rated at GenAI Susceptibility Level 3 (moderate susceptibility), mitigating factors such as basing work on provided images or texts, creating infographics with hashtags, or proposing learning designs based on explanatory videos made it possible for AI to achieve satisfactory results but with more effort. These assignments often incorporated more challenging elements for GenAI, like personal reflections, audio recordings, or custom infographics. Tasks that required the integration of multiple media types or formats also made it more difficult for GenAI tools to generate a complete and coherent response.

In contrast, assignments rated at GenAI Susceptibility Level 4 (high susceptibility) were easier for GenAI tools to handle, though they showed some mitigating factors. For example, assignments often required Word or PowerPoint formats, but incorporated elements like selecting specific works, requiring museum visits, or including images, all of which made them slightly less susceptible. Collaborative and interactive tasks were also present, though not all had robust evaluation criteria to ensure meaningful participation.

Tasks, rated at Level 5 (very high susceptibility), were predominantly in Word format or forum posts and were mostly individual in nature, with only a few interactive or collaborative assignments amongst them. These

tasks required comments, essays, creative proposals, presentations, or analyses, all highly susceptible to AI-generated responses, as they primarily demanded structured, predictable outputs. However, assignments that necessitated creativity, original thought, or specific contextual knowledge, such as discussing works studied in class, consistently showed lower susceptibility. Collaborative or interactive assignments, especially those with clear evaluation criteria for participation, were generally more resistant to GenAI-assisted completion, demonstrating the potential for reducing GenAI susceptibility through thoughtful assessment design.

Phase 2- Expert review and iterative refinement

Drawing on preliminary results, and in consultation with two external GenAI and HE assessment experts, the following criteria were consensually established to be indicative of the operationalisation of the conceptual construct of GenAI susceptibility posed earlier in this paper, as is detailed in Table 4 below:

Table 4. Preliminary criteria and definitions.

Criterion	Definition
Input and Output Format Susceptibility	The ease of transferring assignment instructions and required submission formats to and from GenAI tools.
Intellectual Task Complexity	The degree to which the required cognitive processes align with GenAI capabilities, considering both objective and subjective tasks.
Collaborative Authorship	The extent of genuine collaboration required, beyond mere task division, and its assessment.
Topic Selection	The level of critical thinking required in choosing and justifying the topic.
Opportunity for Creativity	The space provided for original thought that surpasses GenAI-generated patterns.
Instruction Specificity	The level of detail in instructions and its impact on GenAI prompt formulation with greater specificity in instruction leading to greater GenAI susceptibility.
Process-Oriented Assessment	The degree to which the evaluation focuses on the learning process rather than just the final assessment product.
Topic Information Accessibility	The availability of topic-related information to GenAI tools, with more canonical subjects being more susceptible.
Authorship Evidence	The requirement for demonstrable human involvement in the assignment completion.

Consensus diverged around the need to envisage assessment validity and GenAI susceptibility beyond the narrow prism of grading which solely links the issue to student's undeclared GenAI usage, as discussed earlier in the Literature section. Thereby, a streamlined three-point tiered iteration replaced the five-point scale used initially and is to be used in accordance with each of the criteria mentioned in Table 4 previously. Furthermore, a tricolour coding system was also introduced to enhance the visual representation of this division. In an effort to reiterate the designed dialogic and action-oriented nature of the tool, traffic light colours were selected for their connotations with decision-making and intervention. The iterative modifications are summarised in Table 5 below.

Phase 3- End-user validation

Having implemented the expert-informed modifications in Phase 2, this final component sought a return to the practical domain for end-user validation. Short online meetings were held with participants from Phase 1 to present the refined rubric. These sessions guided participants through the application of the new rubric to the assignments submitted previously in Phase 1 and were designed to serve as a forum for garnering feedback on the usability, clarity, and perceived efficacy of the instrument.

Table 5. Summary of post-expert review modifications.

Level	Description	Overall Average	Colour Coding
1- Minimum GenAI Susceptibility	It is very or somewhat difficult to use GenAI tools with satisfactory results.	0-1.5	Green
2- Partial GenAI Susceptibility	It is possible to use GenAI tools with satisfactory results.	1.6-2.5	Amber
3- Total GenAI Susceptibility	It is easy or very easy to use GenAI tools with satisfactory results.	2.6-3	Red

At this stage, of the 34 assessments analysed, a total of 28 assessments were determined to have total GenAI susceptibility, 5 to have partial GenAI susceptibility, and 1 to have minimal GenAI susceptibility overall. Although the same assessments were analysed, on this occasion, the total number of highly susceptible cases rose to 82.35% compared to 67.65% of the total corpus previously with the former iteration. This seems to suggest that the iterative development process has led to a more sensitive instrument for detecting GenAI susceptibilities. Table 6, below, illustrates the cumulative criteria-specific breakdown of the analysis:

Table 6. Criteria-specific breakdown of end-user assessment validity analysis.

Criterion	Level 1	Level 2	Level 3
Input and Output Format	1	5	28
Intellectual Task Complexity	1	4	29
Collaborative Authorship	1	3	30
Topic Selection	1	16	17
Creativity Scope	0	19	15
Instruction Specificity	1	3	30
Process-Oriented Assessment	0	8	26
Topic Information Accessibility	0	4	30
Authorship Evidence	0	4	30

The qualitative suggestions for improvement and usability enhancement were pertaining to minor changes in wording. The resulting definitive iteration of the rubric is presented in the following section.

The PANDORA GenAI Susceptibility Rubric

Having concluded the multistage development process, the novel tool is offered here to spark dialogic deliberation and critical reflection in assessment praxis and policy:

	Level 1 Minimum GenAI Susceptibility *Difficult or very difficult AI use with satisfactory results*	Level 2 Moderate GenAI Susceptibility *Possible AI use with satisfactory results*	Level 3 High GenAI Susceptibility *Easy AI use with satisfactory results*
Input and Output Format	Input/output formats resistant to free GenAI tools: complex multimedia, specialised file types, or formats requiring paid AI services or significant human processing.	Input/output formats with partial GenAI resistance: require human processing of non-text elements; AI output needs conversion to required format.	Input/output formats highly susceptible to GenAI: easily 'pasteable' instructions; submission format directly producible by free AI tools.
Intellectual Task Complexity	Task requires subjective opinion, justification, or complex metacognitive/emotional processes beyond AI capabilities.	Task partially achievable by AI tools but requires some human input.	Task fully achievable by AI tools (e.g., describe, explain, argue, create, or compare within given parameters).
Collaborative Authorship	Genuinely interdependent collaboration required; collaboration explicitly assessed.	Partial collaboration or genuine collaboration not explicitly assessed.	Individual task or easily divisible group work without true collaboration.
Topic Selection	Justified topic choice.	Topic choice without justification.	Topic imposed or limited options without justification.
Creativity Scope	Ample scope for creativity.	Moderate scope for creativity.	Highly limited scope for creativity.

Instruction Specificity	General guidelines only; students must determine specific parameters and justify choices, requiring significant human input for GenAI prompts.	Mix of general and specific instructions; some parameter choices allowed but not justified; minimal student input needed for GenAI prompts.	Highly detailed instructions (e.g., templates, checklists) directly usable as GenAI prompts.
Process-Oriented Assessment	Assessment covers output and developmental process; all steps must be demonstrated	Primary focus on output; some demonstration of developmental process steps required.	Focus solely on output; no developmental process demonstration required.
Topic Information Accessibility	Topic information not readily available online; inaccessible to Gen AI tools.	Topic information available online but difficult to find; partial GenAI access.	Topic information easily accessible online; full GenAI access-
Authorship Evidence	Robust authorship evidence required (e.g., shared online document, explanatory video appearance).	Authorship evidence required but easily manipulated (e.g., editable images).	No authorship evidence required.

The development of the PANDORA GenAI Susceptibility Rubric has brought deeper and more structured insights into assessment design that extend beyond superficial adjustments, revealing deeper complexities in how educational tasks can be manipulated by or remain resilient to generative AI (GenAI) intervention. Initially, group activities were thought to provide a buffer against GenAI misuse under the assumption that collaboration would naturally deter its application. However, further analysis demonstrated that dividing tasks among students did little to reduce susceptibility to GenAI-generated responses. Instead, the extent of genuine collaboration emerged as a pivotal factor. This led to establishing "Collaborative Authorship" as a central criterion within the rubric, rather than relying on mechanical task division, collaborative authorship calls for interactive human engagement, ensuring that the co-construction of knowledge involves critical discussion and shared intellectual responsibility. This key distinction reinforces the need for assessments that reflect collective intellectual rigour rather than piecemeal contributions, where GenAI could easily fill the gaps without meaningful student input.

Additionally, the rubric's development underscored the significance of process-oriented assessment, which evaluates the final product and the steps leading up to it. In doing so, the rubric creates a framework that captures the quality of student collaboration and generates a structured learning pathway that is inherently less vulnerable to GenAI infiltration. This approach ensures that the integrity of the learning process is maintained, as the evaluation hinges on both the process and product, making it increasingly difficult for GenAI to replicate the human-centred aspects of authentic academic work.

In refining the rubric, further attention was given to the intricate relationships between topic selection, creativity, and instruction specificity related to GenAI susceptibility. It became evident that assignments which offered predetermined topics without requiring justification inadvertently opened the door for GenAI misuse. When students are free to select and rationalise their topics, they must engage in critical thinking, directly reducing the rubric's susceptibility criteria. This shift from passive to active intellectual engagement marks a critical pivot in how we conceptualise student assessment in the era of AI.

Moreover, the rubric elevates the importance of creativity and originality in assessments, as these factors challenge the algorithmic and patterned nature of GenAI outputs. Encouraging originality creates a barrier that GenAI, often reliant on formulaic outputs, struggles to overcome. Yet, the rubric also recognises a paradox: overly prescriptive instructions can simplify the process of GenAI prompt formulation, thereby increasing susceptibility. This insight led to the inclusion of the criterion "Instruction Specificity", reminding educators of the delicate balance between clear guidance and fostering student-driven exploration. By offering a more open-ended structure, the rubric promotes student autonomy and ensures that the intellectual demands of the task cannot be easily outsourced to GenAI tools.

Finally, the rubric integrates considerations of topic canonicity and authorship evidence, acknowledging that less conventional subjects and the need for demonstrable human involvement significantly reduce the potential for GenAI manipulation. The ability to differentiate between AI-generated and human-generated work is crucial in maintaining academic integrity, and this rubric provides a structured, methodical approach to achieving that differentiation. In doing so, it serves as a practical tool for educators and an instrument for advancing academic discourse on the evolving challenges that GenAI presents to higher education assessment.

In essence, the PANDORA GenAI Susceptibility Rubric positions itself as a critical tool for the academic community, offering a robust framework that moves beyond reactive measures. It invites educators to engage with the complexities of GenAI in assessment actively, fostering a culture of reflection, innovation, and future-proofed assessment design. Through its multifaceted criteria, the rubric safeguards against GenAI misuse and is a catalyst for more profound, more meaningful learning experiences, thereby reinforcing the educational value of assessments in an AI-driven world. Its importance to the academic world cannot be overstated, as it addresses the urgent need for structured, ethical, and pedagogically sound responses to the disruptive capabilities of GenAI technologies.

Implications for HE praxis and policy

The findings of this study, particularly the development and application of the PANDORA GenAI Susceptibility Rubric, have far-reaching implications for HE policy and practice in the context of increasingly sophisticated GenAI technologies. Firstly, in practice, the novel conceptual definition and the tool offer a structured framework for educators to evaluate and refine their assessment designs. These move beyond traditional moralistic notions of academic integrity towards a more nuanced understanding of assessment validity. Ultimately, this tool may serve to address the issues highlighted in Phase 1 when participants were interviewed. It may thus raise critical awareness amongst educators globally on the realities of progressively capable GenAI tools beyond the GenAI susceptibilities of their present assessments and the possibilities to enhance assessment procedures with and without GenAI tools. Practical implications might also include using the tool to spark institutional dialogue around the need for professional development programmes that equip educators with the skills to design GenAI-resistant and -integrated assessments, fostering collaboration between academic staff, learning designers, and AI specialists to create innovative assessment strategies. Moreover, it is

apparent that making progress on ensuring assessment validity is a complex task. To that end, the authors foresee the practical utility of this tool used as a precursor to build on parallel frameworks such as that crafted by Zaphir et al. (2024). It could furthermore be used to define assessment validity and GenAI susceptibility. Thereafter, frameworks such as Perkins et al.'s (2024) AI Assessment Scale could be used to support the subsequent development process of new assessment offerings based on the findings of the PANDORA GenAI Susceptibility Rubric.

Secondly, at the policy level, our research underscores the urgent need for institutions to move beyond piecemeal approaches and critical silences (Luo, 2024) towards comprehensive, forward-thinking policies that address the multifaceted impacts of GenAI on academic integrity, assessment design, and pedagogical practices. Drawing on the transformative paradigm shift proposed by Dawson et al. (2024), we argue that policy frameworks should prioritise assessment validity over punitive measures, fostering a culture of ethical AI use whilst simultaneously reimagining the purpose and design of assessments in HE. This approach necessitates a reconceptualisation of quality assurance mechanisms (Ryan, 2015; Rawas, 2023) to incorporate GenAI considerations, ensuring that institutional policies are not only reactive but proactive in addressing the challenges and opportunities presented by these technologies.

Drawing upon the broad conceptual underpinnings of the PANDORA GenAI Susceptibility Rubric, its applicability across diverse higher education contexts warrants careful consideration through multiple intersecting dimensions. The rubric's foundational architecture, whilst initially validated within Arts and Humanities, potentially offers considerable adaptability across disciplinary boundaries through its core evaluative criteria. For instance, the 'Intellectual Task Complexity' and 'Process-Oriented Assessment' dimensions might readily translate to STEM disciplines where laboratory work and empirical investigation predominate, whilst 'Collaborative Authorship' and 'Topic Information Accessibility' maintain relevance in professional programmes such as medicine, law, and business studies. The rubric's implementation framework can be contextually calibrated to accommodate varying institutional typologies-—from research-intensive universities to professionally-oriented colleges—-through thoughtful consideration of local assessment policies, quality assurance mechanisms, and pedagogical approaches. Moreover, its adaptability extends to diverse learning modalities, encompassing traditional face-to-face instruction, distance education, and hybrid delivery models, each presenting unique challenges and opportunities for assessment validity. This versatility is particularly salient when considering international and cross-cultural applications, where the rubric's emphasis on authentic assessment and academic integrity can be meaningfully aligned with different national quality frameworks and cultural approaches to knowledge demonstration. The instrument is novel in that it offers a robust yet flexible comprehensive framework that can be systematically implemented across the broader higher education landscape, whilst maintaining sensitivity to discipline-specific requirements, institutional contexts, and diverse student populations.

Limitations

While the instrument offers a novel approach to assessing GenAI susceptibility in HE assessments, several limitations of this study warrant consideration. Firstly, the sample size of 12 lecturers from a single institution in Spain limits the generalisability of the findings. While providing depth, the focus on Arts and Humanities disciplines may not fully capture the nuances of GenAI susceptibility across other academic fields. Additionally, the rapid evolution of GenAI technologies means that the rubric may require frequent updates to remain relevant. The study's reliance on self-reported data from educators could introduce potential biases, particularly given the varying levels of GenAI familiarity among participants. Furthermore, using only two GenAI tools (ChatGPT-3.5 and Claude 3 Haiku) for testing assessment susceptibility may not comprehensively represent the full spectrum of available GenAI technologies. Lastly, the study's context within a Spanish university system may limit its direct applicability to other cultural and educational contexts. Despite these limitations, the PANDORA GenAI Susceptibility Rubric provides a valuable starting point for addressing the complex challenges of GenAI in HE assessment, and future research can build upon this foundation to address these constraints.

Future research directions

Owing to the progressively intricate multi-faceted impacts of human interactions with GenAI technologies in HE, to quantify avenues for future scholarly exploration as multiple may be somewhat of an understatement. Nonetheless, the most evident routes to pursue include the empirical validation in practice in domains cognate to the Arts and Humanities pedagogical setting in which the present study is grounded.

Additionally, future research should explore the applicability of the PANDORA GenAI Susceptibility Rubric across diverse disciplinary contexts, investigate its long-term impact on assessment design and academic integrity, and examine how it might be adapted to keep pace with rapidly evolving GenAI capabilities. Furthermore, studies on the integration of this tool into broader institutional policies and its potential to foster a culture of ethical AI use in academia would be valuable. Lastly, comparative analyses of how different global regions and educational systems approach GenAI-related assessment challenges could provide crucial insights for developing more robust and culturally responsive frameworks.

Conclusion

Amidst the GenAI global regulatory vacuum in HE catalysed through sweeping ambiguities and institutional inertia, this study set out to heed the bold call of Dawson et al. (2024) for a paradigmatic shift in academic integrity that assessment validity matters more than the phenomenon of cheating itself. To that end, the authors sought to conceptually substantiate the deeper epistemological, ontological, and axiological implications of undeclared GenAI use by any

member of the academic community and furthermore produce a novel rubric to facilitate an initial exercise of analysis of current assessment practices. This is by no means a simple task for busy academics in differing pedagogical settings worldwide. To that end, we have been motivated to bridge the oft-muted chasm between research and practice through the development of an interactive digital interface of the rubric, which has been made freely accessible through our dedicated project website. The PANDORA GenAI Susceptibility Rubric is available to interact with here: https://www.pandoraunir.com/

The digitalised version of the tool features an intuitive interface where educators can systematically evaluate their assessments across the criteria through a series of targeted prompts and reflective questions. The platform then generates an analysis report highlighting areas of GenAI susceptibility whilst offering concrete suggestions for assessment redesign tailored to the specific discipline and level of study. This practical resource builds upon our empirical findings created to disseminate a dynamic, evidence-based framework for educators that responds to the pressing needs of time-constrained academics across the sector.

The tool is intended to serve as a dialogical conduit which invites focused critical reflection and subsequent enactment of assessment innovation to not only answer the pressing needs of GenAI susceptibility, but also longstanding documented calls to explore more valid alternatives to traditional approaches such as the essay. As we conclude here, we would like to emphasise our belief that the true power of the PANDORA GenAI Susceptibility Rubric lies not just in its immediate application, but in its potential to spark a broader conversation about the nature of knowledge, learning, and assessment in the digital age. As we stand at this critical juncture in the evolution of HE, we challenge the academic community to use this tool as a springboard for radical innovation, reimagining not just how we assess, but how we teach, learn, and create knowledge in a world where human and artificial intelligence increasingly intertwine.

Acknowledgements

This publication has been funded by the Research Project (PANDORA: Project of Analysis and Development for the Optimisation of Assessment and Regulation of Generative Artificial Intelligence) with reference (PP-2023-02) awarded in the 2023 UNIR Institutional Research Projects call of the Universidad Internacional de La Rioja (UNIR).

References

Ahsan, K., Akbar, S., & Kam, B. (2021). Contract cheating in higher education: A systematic literature review and future research agenda. *Assessment & Evaluation in Higher Education, 47*(4), 1–17. https://doi.org/10.1080/02602938.2021.1931660

Ajjawi, R., Tai, J., Huu Nghia, T. L., Boud, D., Johnson, L., & Patrick, C.-J. (2019). Aligning assessment with the needs of work-integrated learning: The challenges of authentic assessment in a complex context. *Assessment & Evaluation in Higher Education, 45*(2), 1–13. https://doi.org/10.1080/02602938.2019.1639613

Alkaissi, H., & McFarlane, S. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus, 15*(2), e35179. https://doi.org/10.7759/cureus.35179

Allen, S., & Knight, J. (2009). A method for collaboratively developing and validating a rubric. *International Journal for the Scholarship of Teaching and Learning, 3*(2), 10. https://doi.org/10.20429/ijsotl.2009.030210

Ashwin, P. (2020). *Transforming university education: A manifesto*. Bloomsbury Publishing.

Ashwin, P., Goldschneider, B., Agrawal, A., & Smit, R. (2023). Beyond the dichotomy of students-as-consumers and personal transformation: What students want from their degrees and their engagement with knowledge. *Studies in Higher Education, 49*(8), 1439-1450. https://doi.org/10.1080/03075079.2023.2267589

Asiyai, R. I. (2020). Best practices for quality assurance in higher education: Implications for educational administration. *International Journal of Leadership in Education, 25*(5), 843-854. https://doi.org/10.1080/13603124.2019.1710569

Bannister, P. (2024). English medium instruction educator language assessment literacy and the test of generative AI in online higher education. *Journal of Research in Applied Linguistics, 15*(2), 55-72. https://doi.org/10.22055/rals.2024.45862.3214

Bannister, P., Alcalde Peñalver, E., & Santamaría Urbieta, A. (2024a). Transnational higher education cultures and generative AI: A nominal group study for policy development in English medium instruction. *Journal for Multicultural Education, 18*(1/2), 173-191. https://doi.org/10.1108/JME-10-2023-0102

Bannister, P., Alcalde Peñalver, E., & Santamaría Urbieta, A. (2024b). International students and generative artificial intelligence. A cross-cultural exploration of HE academic integrity policy. *Journal of International Students, 14*(3), 149-170. https://doi.org/10.32674/jis.v14i3.6277

Bannister, P., & Carver, M. (2024). 'I don't need professional development; I want institutional development': Legitimising marginalised epistemic capital that disrupts generative AI discourse. *Professional Development in Education*, 1-19. https://doi.org/10.1080/19415257.2024.2427873

Bannister, P., Santamaría Urbieta, A., & Alcalde Peñalver, E. (2023). A Delphi study on generative artificial intelligence and English medium instruction assessment: Implications for social justice. *Iranian Journal of Language Teaching Research, 11*(3), 53-80. https://doi.org/10.30466/ijltr.2023.121406

Baron, P. (2024). Are AI detection and plagiarism similarity scores worthwhile in the age of ChatGPT and other generative AI? *Scholarship of Teaching and Learning in the*

South, 8(2), 151-179. https://doi.org/10.36615/sotls.v8i2.411

Bearman, M., & Ajjawi, R. (2019). Can a rubric do more than be transparent? Invitation as a new metaphor for assessment criteria. *Studies in Higher Education, 46*(2), 1–10. https://doi.org/10.1080/03075079.2019.1637842

Bearman, M., & Luckin, R. (2020). Preparing university assessment for a world with AI: Tasks for human intelligence. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining university assessment in a digital world* (pp. 49-63). Springer.

Bennett, C. (2016). Assessment rubrics: Thinking inside the boxes. *Learning and Teaching, 9*(1), 50-72. https://doi.org/10.3167/latiss.2016.090104

Blackie, M. A. (2024). ChatGPT is a game changer: Detection and eradication is not the way forward. *Teaching in Higher Education, 29*(4), 1109-1116. https://doi.org/10.1080/13562517.2023.2300951

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*(6), 655–670. https://doi.org/10.1080/03075071003777716

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education, 21*(1). https://doi.org/10.1186/s41239-023-00436-z

Boud, D., & Falchikov, N. (2007). *Rethinking assessment in higher education: Learning for the longer term*. Routledge.

Boyd, P., & Hill, J. (2024). Working with rubrics. In N. Reiman, I. Sadler, & J. Hill (Eds.), *Academic standards in higher education. Critical perspectives and practical strategies* (pp. 157-170). Routledge.

Bozkurt, A. (2023). Unleashing the potential of generative AI, conversational agents and chatbots in educational praxis: A systematic review and bibliometric analysis of GenAI in education. *Open Praxis, 15*(4), 261-270. https://doi.org/10.55982/openpraxis.15.4.609

Braun, V., Clarke, V., Hayfield, N., & Terry, G. (2025). Thematic analysis. In P. Brough (Ed.), *Advanced research methods for Applied psychology: Design, analysis and reporting* (pp. 238-248). Routledge.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*(22). https://doi.org/10.3389/feduc.2018.00022

Cash, P., Isaksson, O., Maier, A., & Summers, J. (2022). Sampling in design research: Eight key considerations. *Design Studies, 78*(1), 101077. https://doi.org/10.1016/j.destud.2021.101077

Chaka, C. (2024). Accuracy pecking order – how 30 AI detectors stack up in detecting generative artificial intelligence content in university English L1 and English L2 student essays. *Journal of Applied Learning & Teaching, 7*(1), 127-139. https://doi.org/10.37074/jalt.2024.7.1.33

Chapelle, C. A., & Lee, H. (2021). Conceptions of validity. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 21-34). Routledge.

Condette, M. (2024). 'It's paradoxical, but it works' – towards ambidexterity in external quality assurance: The case of Roman Catholic ecclesiastical higher education. *Studies in Higher Education*, 1-13. https://doi.org/10.1080/03075079.2024.2407516

Dai, Y., Lai, S., Lim, C. P., & Liu, A. (2024). University policies on generative AI in Asia: Promising practices, gaps, and future directions. *Journal of Asian Public Policy,* 1-22. https://doi.org/10.1080/17516234.2024.2379070

Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education*, 1-12. https://doi.org/10.1080/02602938.2024.2386662

De Maio, C. (2024). Institutional responses to ChatGPT: Analysing the academic integrity policies of four public and private institutions of higher education in Australia. *Journal of Academic Language and Learning, 18*(1), T1-T8. https://journal.aall.org.au/index.php/jall/article/view/917

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). SAGE Publications.

Eaton, S. E. (2017). Comparative analysis of institutional policy definitions of plagiarism: A pan-Canadian university study. *Interchange, 48*(3), 271–281. https://doi.org/10.1007/s10780-017-9300-7

Egloff, J. (2024). The college essay is not dead: Using scaffolding and presentations to create ChatGPT-resistant research projects. *Proceedings of the H-Net Teaching Conference, 2*(1), 72–100. https://doi.org/10.33823/phtc.v2i1.225

Ehlers, U. D., & Eigbrecht, L. (2024). *Creating the university of the future. A global view on future skills and future higher education*. Springer.

Farrelly, T., & Baker, N. (2023). Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences, 13*(11), 1109. https://doi.org/10.3390/educsci13111109

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S. D., Köller, O., & Möller, J. (2024). Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. *Computers and Education: Artificial Intelligence, 6,* 100209. https://doi.org/10.1016/j.caeai.2024.100209

Fraile, J., Gil-Izquierdo, M., & Medina-Moral, E. (2023). The impact of rubrics and scripts on self-regulation, self-efficacy

and performance in collaborative problem-solving tasks. *Assessment & Evaluation in Higher Education, 48*(8), 1223–1239. https://doi.org/10.1080/02602938.2023.2236335

Francis, J. E. (2018). Linking rubrics and academic performance: An engagement theory perspective. *Journal of University Teaching & Learning Practice, 15*(1), 1-20. https://doi.org/10.53761/1.15.1.3

Giray, L. (2024). Negative effects of generative AI on researchers: Publishing addiction, Dunning-Kruger effect and skill erosion. *Journal of Applied Learning & Teaching, 7*(2), 398-405. https://doi.org/10.37074/jalt.2024.7.2.38

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). *AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays*. arXiv. https://doi.org/10.48550/arxiv.2304.14276

Kaiser, F., Melo, A. I., & Hou, A. Y. (2022). Are quality assurance and rankings useful tools to measure 'new' policy issues in higher education? The practices in Europe and Asia. *European Journal of Higher Education, 12*(1), 391-415. https://doi.org/10.1080/21568235.2022.2094816

Kaqinari, T. (2023). Facilitators and barriers to online teaching and educational technology use by university lecturers during COVID-19: A systematic review of qualitative evidence. *Trends in Higher Education, 2*(4), 636–666. https://doi.org/10.3390/higheredu2040038

Kenworthy, A. L., & Hrivnak, G. A. (2014). To rubric or not to rubric. *Journal of Management Education, 38*(3), 345–351. https://doi.org/10.1177/1052562914530103

Kirmani, A. R. (2023). Fantastic (AI) bots and how to catch them. *MRS Bulletin, 48*(4), 314-318. https://doi.org/10.1557/s43577-023-00516-5

Knight, S., Dickson-Deane, C., Heggart, K., Kitto, K., Çetindamar Kozanoğlu, D., Maher, D., Narayan, B., & Zarrabi, F. (2023). Generative AI in the Australian education system: An open data set of stakeholder recommendations and emerging analysis from a public inquiry. *Australasian Journal of Educational Technology, 39*(5), 101-124. https://doi.org/10.14742/ajet.8922

Kohnke, L., Zou, D., & Moorehouse, B. L. (2024). Technostress and English language teaching in the age of generative AI. *Educational Technology & Society, 27*(2), 306-320. https://doi.org/10.30191/ETS.202404_27(2).TP02

Kramm, N., & McKenna, S. (2023). AI amplifies the tough question: What is higher education really for? *Teaching in Higher Education, 28*(8), 2173-2178. https://doi.org/10.1080/13562517.2023.2263839

Krebs, R., Rothstein, B., & Roelle, J. (2022). Rubrics enhance accuracy and reduce cognitive load in self-assessment. *Metacognition and Learning, 17*(2), 627-650. https://doi.org/10.1007/s11409-022-09302-1

Kumar, R., & Mindzak, M. (2024). Who wrote this? *Canadian Perspectives on Academic Integrity, 7*(1), 1-9. https://doi.org/10.55016/ojs/cpai.v7i1.77675

Lalor, J., Lorenzi, F., & Rami, J. (2015). Developing professional competence through assessment: Constructivist and reflective practice in teacher-training. *Eurasian Journal of Educational Research, 15*(58), 45-66 https://dx.doi.org/10.14689/ejer.2015.58.6

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 6,* 100174. https://doi.org/10.1016/j.caeo.2024.100174

Leaver, T., & Srdarov, S. (2023). ChatGPT isn't magic: The hype and hypocrisy of generative artificial intelligence (AI) rhetoric. *M/C Journal, 26*(5). https://doi.org/10.5204/mcj.3004

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Lipnevich, A. A., Guskey, T. R., Murano, D. M., & Smith, J. K. (2020). What do grades mean? Variation in grading criteria in American college and university courses. *Assessment in Education: Principles, Policy & Practice, 27*(5), 480–500. https://doi.org/10.1080/0969594x.2020.1799190

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., & Hu, H. (2023). *ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models*. arXiv. https://doi.org/10.48550/arXiv.2304.07666

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Lodge, J. M., Yang, S., Furze, L., & Dawson, P. (2023). It's not like a calculator, so what is the relationship between learners and generative artificial intelligence? *Learning: Research and Practice, 9*(2), 117-124. https://doi.org/10.1080/23735082.2023.2261106

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*, 1–14. https://doi.org/10.1080/02602938.2024.2309963

Matthews, J. A., & Volpe, C. R. (2023). Academics' perceptions of ChatGPT-generated written outputs: A practical application of Turing's imitation game. *Australasian Journal of Educational Technology, 39*(5), 82-100. https://doi.org/10.14742/ajet.8896

McDonald, N., Johri, A., Ali, A., & Hingle, A. (2024). *Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines.* arXiv. https://doi.org/10.48550/arXiv.2402.01659

McGraw, A., & Mason, M. (2021). The formulaic essay and

its power to de-fuse reading responses. *Literacy Learning: The Middle Years, 29*(3), 8-15. https://search.informit.org/doi/10.3316/informit.069867974654568

Melekhina, E. A., & Levitan, K. M. (2015). Assessment system in writing essays by graduate students. *Procedia - Social and Behavioral Sciences, 200*, 482–489. https://doi.org/10.1016/j.sbspro.2015.08.099

Meylani, R. (2024). A comparative analysis of traditional and modern approaches to assessment and evaluation in education. *The Western Anatolian Journal of Educational Sciences, 15*(1), 520–555. https://doi.org/10.51460/baebd.1386737

Mirador, J. (2018). "You need to criticize, not just summarize!" Investigating "criticality" in written assignments of postgraduate students. *Writing & Pedagogy, 10*(1-2), 61–92. https://doi.org/10.1558/wap.33366

Möck, L. A. (2022). Prediction promises: Towards a metaphorology of artificial intelligence. *Journal of Aesthetics and Phenomenology, 9*(2), 119-139. https://doi.org/10.1080/20539320.2022.2143654

Monett, D., & Grigorescu, B. (2024). *Deconstructing the AI myth: Fallacies and harms of algorithmification*. ResearchGate. https://www.researchgate.net/publication/382802495

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education, 9*(3), e3292. https://doi.org/10.1002/rev3.3292

Morton, J. K., Northcote, M., Kilgour, P., & Jackson, W. A. (2021). Sharing the construction of assessment rubrics with students: A model for collaborative rubric construction. *Journal of University Teaching & Learning Practice, 18*(4), 1-17. https://doi.org/10.53761/1.18.4.9

Newton, P., & Xiromeriti, M. (2024). ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assessment & Evaluation in Higher Education, 49*(6), 781-798. https://doi.org/10.1080/02602938.2023.2299059

Nguyen, A., Hong, Y., Dang, B., & Huang, X. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education, 49*(5), 847-864. https://doi.org/10.1080/03075079.2024.2323593

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*, 100329. https://doi.org/10.1016/j.edurev.2020.100329

Panadero, E., Jönsson, A. Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review, 35*(4), 113. https://doi.org/10.1007/s10648-023-09823-4

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 and Learning Practice, 21*(6), 1-18. https://doi.org/10.53761/q3azde36

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education, 21*(1), 1-25. https://doi.org/10.1186/s41239-024-00487-w

Perris, K., & Mohee, R. (2020). *Quality assurance rubric for blended learning*. Commonwealth of Learning. https://oasis.col.org/items/2d37644b-4134-4f91-807d-e48ab8d26557

Pitt, P., Dullaghan, K., & Sutherland-Smith, W. (2020). "Mess, stress and trauma": Students' experiences of formal contract cheating processes. *Assessment & Evaluation in Higher Education, 46*(4), 1–14. https://doi.org/10.1080/02602938.2020.1787332

Pratschke, B. M. (2024). *Generative AI and education: Digital pedagogies, teaching innovation and learning design*. Springer.

Ragupathi, K., & Lee, A. (2020). Beyond fairness and consistency in grading: The role of rubrics in higher education. In C. S. Sanger, & N. W. Gleason (Eds.), *Diversity and inclusion in global higher education. Lessons learnt from Asia* (pp. 73-96). Palgrave Macmillan.

Rawas, S. (2023). ChatGPT: Empowering lifelong learning in the digital age of higher education. *Education and Information Technologies, 29*(6), 6895-6908. https://doi.org/10.1007/s10639-023-12114-8

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching, 6*, 342-363.https://doi.org/10.37074/jalt.2023.6.1.9

Ryan, T. (2015). Quality assurance in higher education: A review of literature. *Higher Learning Research Communications, 5*(4), 1-12. https://doi.org/10.18870/hlrc.v5i4.257

Sadler, I., Reimann, N., & Sambell, K. (2022). Feedforward practices: A systematic review of the literature. *Assessment & Evaluation in Higher Education, 48*(3), 1–16. https://doi.org/10.1080/02602938.2022.2073434

Sharples, M. (2022). Automated essay writing: An AIED opinion. *International Journal of Artificial Intelligence in Education, 32*(4), 1119-1126. https://doi.org/10.1007/s40593-022-00300-7

Smolansky, A., Cram, A., Raduescu, C., Zeivots, S., Huber, E., & Kizilcec, R. F. (2023). Educator and student perspectives on the impact of generative AI on assessments in higher education. *Proceedings of the Tenth ACM Conference on Learning @ Scale*. https://doi.org/10.1145/3573051.3596191

Sok, S., & Heng, K. (2024). *Generative AI in higher education:*

The need to develop or revise academic integrity policies to ensure the ethical use of AI. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4806030

Song, N. (2024). Higher education crisis: Academic misconduct with generative AI. *Journal of Contingencies and Crisis Management, 32*(1), e12532. https://doi.org/10.1111/1468-5973.12532

Sweeney, S. (2023). Who wrote this? Essay mills and assessment – considerations regarding contract cheating and AI in higher education. *The International Journal of Management Education, 21*(2), 100818. https://doi.org/10.1016/j.ijme.2023.100818

Tafazoli, D. (2024). Exploring the potential of generative AI in democratizing English language education. *Computers and Education: Artificial Intelligence, 7*, 100275. https://doi.org/10.1016/j.caeai.2024.100275

Tan, K. H. K. (2020). Rubrics, power and conduct. In K. H. K. Tan (Ed.), *Assessment rubrics decoded* (pp. 79-91). Routledge.

Tang, K., Cooper, G., Rappa, N., Cooper, M., Sims, C., & Nonis, K. (2024). A dialogic approach to transform teaching, learning & assessment with generative AI in secondary education: A proof of concept. *Pedagogies: An International Journal, 19*(3), 493-503. https://doi.org/10.1080/1554480x.2024.2379774

Taylor, S. (2024). *Is UNESCO's guidance for Generative AI (GenAI) in education & research "the solution" to the problem of GenAI adaptation in international schools? A critical analysis of policy recommendations through Carol Bacchi's What's The Problem Represented to Be? (WPR) Framework. Wayfinder Learning Lab.* https://sjtylr.net/wp-content/uploads/2024/09/taylors_edd_-policy_unescogenai.pdf

Teixeira, R. J., Brandão, T., & Dores, A. R. (2022). Academic stress, coping, emotion regulation, affect and psychosomatic symptoms in higher education. *Current Psychology, 41,* 7618–7627. https://doi.org/10.1007/s12144-020-01304-z

Tight, M. (2024). Challenging cheating in higher education: A review of research and practice. *Assessment & Evaluation in Higher Education,* 1–13. https://doi.org/10.1080/02602938.2023.2300104

UNESCO. (2023). *Guidance for generative AI in education and research.* https://unesdoc.unesco.org/ark:/48223/pf0000386693

Vallis, C. (2024). Authentic assessment in higher education: The spectre of lost futures. *Teaching in Higher Education*, 1-8. https://doi.org/10.1080/13562517.2024.2362217

Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science, 7*(1), 20220158. https://doi.org/10.1515/opis-2022-0158

Waltzer, T., Pilegard, C., & Heyman, G. D. (2024). Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity, 20*(1), 11. https://doi.org/10.1007/s40979-024-00158-3

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*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

West, H., Malcolm, G., Keywood, S., & Hill, J. (2019). Writing a successful essay. *Journal of Geography in Higher Education, 43*(4), 609–617. https://doi.org/10.1080/03098265.2019.1655720

Wilkerson, J. R. (2020). Rubrics meeting quality assurance and improvement needs in the accreditation context. *Quality Assurance in Education, 28*(1), 19–32. https://doi.org/10.1108/QAE-04-2019-0045

Zaphir, L., Lodge, J. M., Lisec, J., McGrath, D., & Khosravi, H. (2024). *How critically can an AI think? A framework for evaluating the quality of thinking of generative artificial intelligence*. arXiv. https://doi.org/10.48550/arXiv.2406.14769

Zemits, B. I. (2017). Representing knowledge: Assessment of creativity in humanities. *Arts and Humanities in Higher Education, 16*(2), 173–187. https://doi.org/10.1177/1474022215601862