Content list available at http://ijltr.urmia.ac.ir

Iranian Journal of Language Teaching Research

ORIGINAL ARTICLE

Urmia University

A Delphi Study on Generative Artificial Intelligence and English Medium Instruction Assessment: Implications for Social Justice

Peter Bannister [a, *], Alexandra Santamaría-Urbieta [a], Elena Alcalde-Peñalver [b]

[a] *Universidad Internacional de La Rioja (UNIR), Spain*

[b] *Universidad de Alcalá, Spain*

ABSTRACT

The emergence of generative artificial intelligence (GenAI) text generator tools and the potential challenges for higher education (HE) have characterised informal academic discussion on multiple fora. Specifically examining the case of English medium instruction (EMI) assessment academic integrity, this study sought to explore this conundrum by conceptualising threats and possible recommendations to counter these by creating a problem-solution matrix for key stakeholders considering the scarce academic literature available. An exploratory Delphi technique was employed as a way of generating ideas, gauging expert perspectives, and establishing consensus based on the premise of wisdom-of-(expert)-crowds. In the data collection stage, this new use of the mixed-methods methodology in the field included iterative Delphi questionnaire rounds and concurrent focus group sessions with a panel of 26 international experts. Quantitative and qualitative data were analysed using descriptive statistics and thematic analysis, respectively. The resulting GenAI and EMI Assessment Problem-Solution Matrix is an empirically informed instrument for key stakeholders in EMI HE that exemplifies a range of GenAI-induced issues and recommendations as to how to proceed going forward in EMI HE pedagogical settings. This contributes to the field in line with broader theoretical assessment principles, particularly with those seeking to mitigate inequitable practices. Further contextual matters pertaining to social justice were highlighted, such as the effects of the massification and commodification of HE on the role of assessment in both EMI didactic contexts and others. The findings here take a step towards addressing the gaps identified but also represent a means of sparking much-needed further discussion in both extant literature and praxis.

Keywords: academic integrity; assessment; English as a medium of instruction; generative artificial intelligence; higher education

ARTICLE HISTORY

Received: 8 July 2023 **Revised version received:** 17 Sept. 2023

Accepted: 11 Oct. 2023 **Available online:** 15 Dec. 2023

* Corresponding author: Universidad Internacional de La Rioja (UNIR), Spain
Email address: peter.bannister@unir.net

10.30466/ijltr.2023.121406

Introduction

Deontology may be defined as an ethical theoretical prism through which one may discern between right and wrong in accordance with a particular set of rules. This premise, prima facie, appears to offer a cornerstone on which academic integrity in higher education (HE) would not only seemingly be sustained but, in fact, prosper. However, the long-standing contrast documented between theory and practice (e.g., Denisova-Schmidt, 2017) would suggest that HE academic integrity preservation is a more testing and, at times, complicated pursuit. Amidst the many associated conundrums, alongside headliners such as contract cheating and plagiarism (Macfarlane, et al., 2012), generative artificial intelligence (GenAI)-assisted academic misconduct has emerged as a burgeoning protagonist that seems to only add further complication to the matter (Morreel et al., 2023).

Recent media reports have brought to light instances of students potentially misusing GenAI tools to shortcut academic work and gain unfair advantages. For example, some students have purportedly used large language models to generate entire research essays by simply prompting GenAI applications with their paper topics and parameters. These tools then produce original-sounding content with appropriate formatting and citations, requiring little effort from the student. Additionally, there are accounts of students potentially employing AI software to automatically answer exam questions in online courses with customised responses aimed at duping automated proctoring systems (Fergus et al., 2023). These instances illustrate how GenAI could be leveraged to allow students to plagiarise, cheat, and skirt academic integrity measures at scale. Such cases underscore the urgent need for research examining GenAI's implications for assessment security, educator preparedness, and policies, which our study aims to address.

In this landscape, despite comprehensive international sector-wide take up and implementation of a range of digital tools to safeguard academic integrity, such as Turnitin, extant literature on these developments attests to a fast-paced and ever-changing conflict (Rudolph et al., 2023) that arguably strains the once unbreakable vow between HE teaching practitioner and technology as a quintessential staple of good pedagogical practice (Cabaleiro-Cerviño & Vera, 2020; Gutiérrez-Martín et al., 2022). Such digital detection instruments may seem to be conducive to effective self-learning and assessment and fruitful peer review (Chew et al., 2015; Li & Li, 2018). However, they are not without their limits in the current panorama. For instance, they might be used as a potentially punitive but questionably effective deterrent instead of an effective tool for developing academic culture awareness (Kaktiņš, 2019; Li & Li, 2018). Furthermore, they could also be seen as a comprehensive and established means of sounding the alarm on concerning AI-assisted academic misconduct at the time of writing (Khalil & Er, in press; Sadasivan et al., 2023).

The most common methods employed in bilingual education are English Medium Instruction (EMI), which is favoured at higher institutions, and Content and Language Integrated Learning (CLIL), which is concentrated in primary and secondary schools (Dearden, 2014). The European push for plurilingualism gave rise to EMI, which has since become a global phenomenon (Dearden, 2014). By employing English to teach academic courses in non-Anglophone situations, Dearden (2014, p. 4) defines EMI as using English to "teach academic subjects in countries or jurisdictions where the first language (L1) of the majority of the population is not English". Many EMI instructors are concerned, however, about the execution of institutional strategies, which are driving EMI's expansion and are intended to draw in more international students and improve university rankings (Lasagabaster, 2018). Some claim that teaching language is distinct from teaching content (Airey, 2012), and there is a lack of clarity on the use of only English vs. multilingual techniques (Dearden, 2014). Owing to this, some EMI teachers resort to trial-and-error methods in their practice while feeling anxious and powerless (Doiz & Lasagabaster, 2018). To make matters worse, as Farrell (2019, p. 278) remarks, "gap exists between the top-down pressure to incorporate EMI programs and the bottom-up EMI teacher implementation of these programs without any real institutional support or clear pedagogical guidelines to follow."

Assessments in EMI contexts have been traditionally considered problematic even before the emergence of advanced GenAI writing tools like ChatGPT (Dearden, 2014). This suggests assessments may be an especially vulnerable area for EMI programmes in light of these latest developments. Language models could significantly impact key pedagogical aspects of EMI, including evaluating language proficiency, developing academic writing, and enabling cross-cultural communication. The linguistic and authorial risks posed by potential GenAI misuse are thus particularly severe threats to academic integrity in EMI higher education (Lasagabaster, 2022). However, at this time, the authors were unable to find scholarly literature or expert consensus examining these risks of generative AI for EMI assessment specifically. This represents a concerning gap given assessments' fundamental role in EMI programmes.

The rise of generative AI technologies prompts timely deliberation regarding their implications for social justice and equity in EMI assessment. As this study explores, advanced generative models like ChatGPT potentially allow students to shortcut academic writing assignments and exams which leads to assessment integrity being undermined. However, these technologies may disproportionately harm specific student groups, and, in turn, exacerbate inequities. EMI assessments, which evaluate content mastery and English proficiency, have been found to create bias against linguistically marginalised students (Milligan & Tikly, 2016; Mortenson, 2022). GenAI could amplify these injustices if used by privileged students to artificially boost academic performance (Dai et al., 2023). The development of proactive policies and pedagogical practices is therefore needed to promote equitable EMI assessment in the age of GenAI (Shamim, 2023). This study thus aims to build consensus on threats and solutions to guide institutions seeking socially just assessments measuring true student abilities, not advantages conferred by unethical AI utilisation. Fostering academic integrity through ethical AI usage is paramount for inclusion and fairness.

In this context, the authors of this study set out to carry out an exploration of the challenges posed by AI text generator tools to academic integrity in HE, with a specific focus on EMI assessment. The objective was to identify potential threats and develop recommendations to address them, ultimately creating a problem-solution matrix based on expert consensus.

Literature Review

Turning to related extant literature, it is highly interesting to note that, despite the resounding mediatic coverage pertaining to GenAI-powered text generator tools and their implications for the very existence of HE as we know it, there is an extremely limited number of scholarly publications available (e.g., Abd-Elaal et al., 2022; Bishop, in press; Lund & Ting, in press; Shen et al., in press), highlighting in the first instance the gaps in formalised scholarly discussion. This would also appear to reflect not only the juvenile status of the area of study but also the contrast between the gradual and time-consuming nature of research and development for publication and that of the swiftly changing landscape that is the object of study. Considering this, applications such as ChatGPT or GPT-4 may be the posterchild at present for AI-powered text generation tools. However, in line with Passey (2019), the authors have made the conscious decision to focus on the underlying theoretical assumptions and principles in play as opposed to focusing the present study on any single given application. This decision was taken to address the issues at the heart of the matter as a means of addressing the theoretical gaps which have been identified thus far.

In contrast, numerous comprehensive and theoretically credible assessment and evaluation alternatives abound in scholarship which eschew the tenets of testing per se in favour of the longer-term development procedures associated with assessment (Lynch, 2001), such as those which follow the models of assessment for and as learning (e.g., Carless et al., 2017 among many others). However, a further gap has been identified; thereby, the bibliographic exploration undertaken at the time of writing was not able to locate literature which examined the potential correlation between such proposals and assessment principles and their validity in the shadow of GenAI.

In the field of digital ethics, regarding academic integrity and digital tool use for academic misconduct there is a general agreement that cheating and dishonesty are pervasive issues in educational institutions, and the rise of digital technologies has made it easier for students to engage in such activities (Crawford et al., 2023). There is growing concern about the use of AI tools can potentially be exploited for academic cheating and intentional dissemination of false information (Chan, 2023). Furthermore, there have been calls to develop ethical principles and guidelines for AI in education to ensure that AI tools are used responsibly and ethically (Lodge et al., 2023). This involves considering the implications of AI technologies for students, teachers, technology developers, policymakers, and institutional decision-makers (Escotet, 2023). That is to say, the importance of academic integrity and the responsible use of digital tools is emphasised for all key stakeholders. Although, it should be noted that the operationalisation of this call requires a proactive approach in addressing academic misconduct both in practice and scholarship, which ought to include the development of ethical guidelines and the use of technology to detect and prevent dishonest behaviour (Mhlanga, 2023). At the time of writing, instances of uptake in practice have been found to be extremely limited.

A further lack of true or knowable answers, such as decision-making, policy, or long-range forecasting was identified in relation to the specific pedagogical setting of focus in which this study is grounded. EMI has become a salient feature of the HE horizon on a global scale, and, unsurprisingly, much research activity has been undertaken to investigate differing issues of note on this complex educational phenomenon (see the seminal work of Macaro et al. (2018) for an in-depth overview). However, even in a more recent update of the previously cited work, Macaro (2022) identifies a significantly finite quantity of EMI assessment publications and thus confirms previous concerns of scarce EMI assessment literature availability by Kao and Tsou (2017). There is however a limited range of relevant works to the focus of this study, such as those written by Inbar-Lourie (2022) and Li and Wu (2018), who address the absence of formative assessment approaches and learning-orientated classroom assessment in EMI contexts, respectively.

These authors identify numerous established challenges in EMI assessment; for example, assessing students in a language they are still learning is a major challenge. Their performance on assessments in English as a second language is likely affected by their English proficiency rather than purely reflecting their understanding of academic content (Robinson, 2010; van der Walt & Kidd, 2013). Therefore, assessment results may not accurately measure students' acquisition of subject knowledge. A potential solution is for EMI teachers to be clear about the focus of assessment - whether on subject knowledge, language skills, or both (Coyle, Hood & Marsh, 2010). However, even greater challenges reside in choosing appropriate assessment methods aligned to instructional objectives and implementing assessments consistent with their purpose. When the goal is for students to demonstrate academic knowledge, efforts should be made to avoid penalising students for their English language abilities. Conversely, if developing academic English skills is an important or relevant objective for a subject course, EMI teachers are recommended to incorporate measures promoting language learning (Tai, 2015).

As can be seen from the years of publication of the cited works, these are longstanding problems in EMI assessment. However, even in Macaro's (2022) updated review, there are no studies which specifically address the implications of GenAI for EMI assessment. On further inspection, to date, no such studies have been localizable. At this point, it may be tempting to speculate on the possible student use of GenAI writing tools to pass off swathes of text produced by ChatGPT and other apps as their own, which clearly undermines the premise of gauging linguistic competency in EMI assessment. There may also be the inclination to wonder about assessment security of non-proctored EMI formative assignments, such as the essay. In this context, it is therefore understandable that the creation of apparently promising GenAI text classifiers was initially enthusiastically received as a deterrent. Although a scant number of studies have begun to emerge proposing other alternatives, such as that penned by Rudolph et al. (2023), who suggest moving towards authentic assessment methods, this work deals with HE assessment generally and does not account for the particularities of EMI HE academic integrity. However, further examination has called into question the validity and effectiveness of such detection programs thus far (e.g., Dalalah & Dalalah, 2023).

Moreover, Liang et al. (2023) raised concerns of bias against submissions produced by learners who use English as an additional language may be particularly vulnerable to the present pitfalls of this technology. This means that students already navigating educational inequities due to their linguistic backgrounds face further marginalization by flawed AI systems. Predictive algorithms reflecting the biases of their training data can penalise non-native stylistic and grammatical quirks, however slight. Consequently, multilingual students may be unfairly accused of AI-generated writing. Such misclassification risks harming vulnerable students' academic trajectories and perpetuating injustices. To promote equity, developing transparent, ethical AI systems that accommodate diverse student populations is therefore a critical task now more than ever. To this end, institutions must partner with linguistically diverse communities to enhance classifiers and mitigate prejudice. Failing to address bias in emergent systems used to uphold academic integrity may further exclude linguistically marginalised groups.

The scarcity of scholarly exploration is also echoed in practice both in the classroom, with teacher practitioner assessment literacy issues highlighted (Otto and Estrada Chichón, 2021), and at an institutional level, with only 22 of a total of 55 countries studied which had made institution-wide EMI-specific policy provision (Dearden, 2014). Even in a more recent review, seldom instances of institutional policy development can be found, and even these seemingly do not comprehensively cater for the pitfalls of EMI assessment internationally (e.g., UCL, 2023). More recent calls from authors such as Hultgren et al. (2022) to tackle the lack of exploration in scholarship and highly limited examples of institutional policy development in practice taking place in the further concerning background of the commodification and massification of HE, which these also denounce, are therefore seemingly yet unanswered. The commodification and massification of higher education, driven by the pursuit of profit and prestige, has led many institutions to prioritize recruiting large numbers of international students without equitable regard for social justice (Milligan & Tikly, 2018). This inequitable approach often fails to provide adequate language and academic support (Mortenson, 2021), creating conditions ripe for GenAI misuse that further disadvantage already marginalised multilingual students in assessments meant to evaluate proficiency. Interestingly, in an EMI research agenda written by Sah (2022), both EMI assessment and issues of social justice within the field are highlighted as key areas in need of further exploration.

In sum, the authors identified several concerning gaps in the scholarly literature regarding the implications of GenAI text generation tools for assessment and academic integrity in EMI HE contexts. Specifically, they note the extremely limited number of published studies examining this issue thus far, highlighting deficiencies in formal academic discussion, likely reflecting the nascent status of this research area. Moreover, there is a general scarcity of assessment literature pertaining to EMI contexts, with very few studies exploring formative assessment, learning-oriented approaches, or the challenges of evaluating EMI students' content knowledge versus language skills. Most critically, no identifiable studies have addressed the implications of GenAI text generation for safeguarding academic integrity and assessment in EMI HE. Relatedly, there is a lack of institutional policies comprehensively dealing with EMI assessment matters considering GenAI developments. Calls to tackle these gaps in research and practice are issued amidst the concerning backdrop of HE commodification and massification yet seem largely unheeded, once again highlighting important social justice concerns. Additional lacunas include the lack of scholarly analysis on aligning assessment proposals to AI environments and achieving expert consensus on this issue specifically for EMI assessment.

These gaps, the lack of documented empirically-informed potential GenAI-related threats posed to academic integrity in EMI settings and bespoke recommendations for EMI assessment motivated the implementation of this study. The lack of a true or knowable answer, such as decision-making, policy, or long-range forecasting in this regard subsequently led to the creation of a GenAI and EMI Assessment Problem-Solution Matrix for key stakeholders in the sector, grounded in the context of HE was proposed by the authors. To that end, having identified the substantive gaps in the scholarly literature and in EMI assessment practice, the following research questions (RQs) were established for exploration in this study:

- RQ1: What expert consensus can be reached on the challenges posed by AI text generators for assessment on EMI programmes of study?

- RQ2: How would experts concordantly operationalise recommendations for EMI assessment design to address GenAI-related threats to academic integrity?

To address the RQs and create the proposed problem-solution matrix whilst considering the status quo in extant literature and assessment praxis, the Delphi method was selected as an established means of generating ideas, gauging expert opinion, and establishing consensus (Linstone & Turoff, 1976). There are, however, several relevant longstanding controversies and challenges associated with the methodology, which the researchers acknowledged. These include the bandwagon effect, vulnerability to manipulation, and reticence of stance modification in the presence of others (Morgan, 1997). These issues are particularly acute in the present study as the expression and modification of opinion throughout are key to constructing and negotiating the knowledge which forms the basis of expert consensus. A potential risk of this may be highlighted in the closing section of the study. Thereby, panellists may simply feel compelled to agree to the final iteration of the matrix (Sterling et al., 2023).

The subjective nature of gathering expert opinions through methods like the Delphi technique raises important questions around validity and reliability. The validity of results can vary substantially based on how experts are selected, prompting critiques regarding reliability across different studies (Greenbaum, 1998). Moreover, heavy reliance on experts' subjective judgements makes the Delphi technique susceptible to preconceived biases (Linstone & Turoff, 1976). Although consensus may be quantified statistically, this can potentially disguise lingering disagreements amongst experts. Thus, while offering benefits, Delphi studies must grapple with challenges stemming from their dependence on subjective expert opinions as primary data sources that may yield different results if the exercise were to be repeated with different participants (Fink-Hafner et al., 2019). Careful consideration of validity, reliability, and steps to minimize bias is required when designing and interpreting Delphi studies. The novel bespoke design used here is detailed in the subsequent section.

Method

Design

The Delphi method, initially developed during the Cold War by Olaf Helmer and Norman Dalkey of the Rand Corporation in the 1950s, is a structured communication technique used for forecasting. Originally designed to predict the impact of technology on warfare, it has since been employed in various fields such as healthcare, education, and public policy-making. Keying into the wisdom-of-(expert)-crowds (Surowiecki, 2005), this method consists in consensus establishment by eliciting and refining expert knowledge through asynchronous rounds of iterative questionnaires (Scheibe et al., 2002). This traditionally involves setting forecasting tasks, collecting initial forecasts and justifications, providing feedback, and repeating the process until a satisfactory level of consensus is achieved. Although the methodology has been widely used in the field of healthcare, there have been calls for wider use and application of the method in other fields (Sterling et al., 2023) to build on limited uptake thus far. For instance, Uztosun (2018) sought to define expert consensus on the necessary professional competences to teach English at primary schools in Turkey. In marginally closer thematic focus to the present study, Chen and Saulter (2019) used the Delphi method to define expert consensus on the specific challenges test developers encounter on the administration of L2 writing assessments to test-takers with different disabilities.

There are advantages to the methodology for this study in that, epistemologically, thanks to expert domain-specific knowledge (Green, 2014), expertise generalisability favourably affords concurrent validity and reliability (Cuhls, 2001) to findings. This methodology is specifically beneficial to the study as it leverages collective expertise on the emerging issues at hand in a structured process aimed at consensus building. This can provide insightful data on expert opinions to inform policies and practices for EMI assessment considering AI advancement and its potential implications for EMI assessment integrity. Thus, this approach was deemed to be particularly justified to address the RQs.

Nonetheless, drawbacks may include a high attrition rate, elevated time investment, and limited informant opinion elaboration opportunities (Chan, 2022). Therefore, the authors of this study have specifically adapted the established methodology by creating synchronous and asynchronous components, as is detailed in Figure 1 below:

In short, this novel adaptation was created with the aim of mitigating the shortcomings highlighted and entailed both asynchronous and synchronous iterative questionnaire rounds with interjacent focus group sessions. This design aimed to allow participants to reflect on and review opinions on the focus of the study whilst also considerably reducing the time investment. Microsoft Forms and Microsoft Teams were employed to carry out the different stages of the investigation.

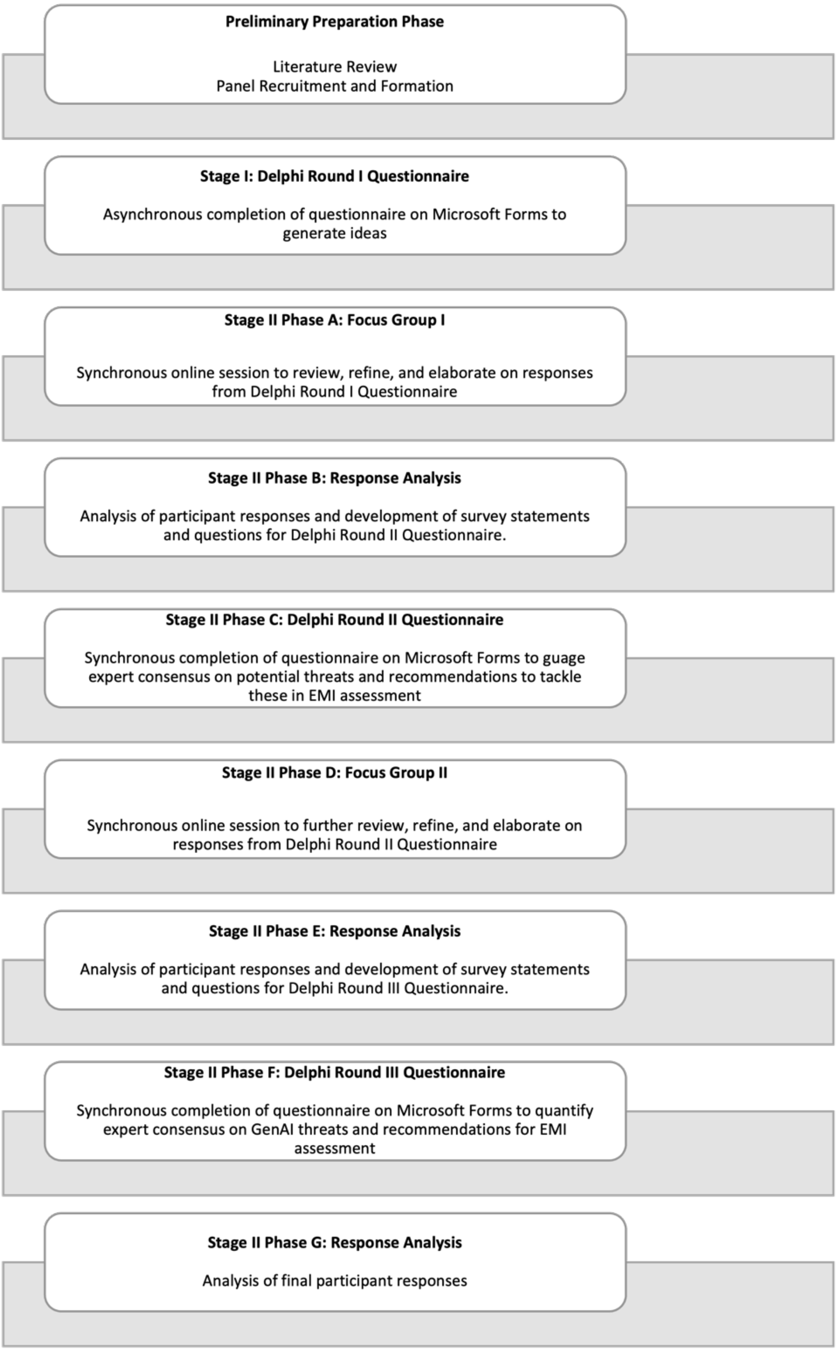
Preliminary Preparation Phase
Literature Review
Panel Recruitment and Formation
Stage I: Delphi Round I Questionnaire
Asynchronous completion of questionnaire on Microsoft Forms to generate ideas
Stage II Phase A: Focus Group I
Synchronous online session to review, refine, and elaborate on responses from Delphi Round I Questionnaire
Stage II Phase B: Response Analysis
Analysis of participant responses and development of survey statements and questions for Delphi Round II Questionnaire.
Stage II Phase C: Delphi Round II Questionnaire
Synchronous completion of questionnaire on Microsoft Forms to guage expert consensus on potential threats and recommendations to tackle these in EMI assessment
Stage II Phase D: Focus Group II
Synchronous online session to further review, refine, and elaborate on responses from Delphi Round II Questionnaire
Stage II Phase E: Response Analysis
Analysis of participant responses and development of survey statements and questions for Delphi Round III Questionnaire.
Stage II Phase F: Delphi Round III Questionnaire
Synchronous completion of questionnaire on Microsoft Forms to quantify expert consensus on GenAI threats and recommendations for EMI assessment
Stage II Phase G: Response Analysis
Analysis of final participant responses

Figure 1. Research Design Structure

Participants

The selection criteria for identifying experts were rigorously determined by the authors, who considered three primary dimensions: knowledge, experience, and pedagogical policy responsibility. The inclusion of knowledge as a criterion here stems from the recognition that experts must possess an extensive comprehension of the fundamental concepts, theories, and skills within their respective domain. This knowledge serves as the cornerstone upon which their expertise is built. The second criterion, experience, was chosen due to its complementary role alongside conceptual knowledge. Experts must have acquired applied experiential knowledge through substantial professional practice to cultivate robust expertise. Thus, the authors actively sought participants with a substantial track record of working in relevant roles. Lastly, pedagogical policy responsibility was identified as a key criterion because genuine experts not only amass knowledge and experience but also occupy positions of influence where they can enact meaningful decisions concerning teaching and training practices. Experts are distinguished not only by their capacity to apply their expertise but also by their ability to translate it into impactful policies and initiatives. The full range of inclusion and exclusion criteria feature in Table 1 below:

Table 1
Summary of Inclusion and Exclusion Criteria

	Inclusion criteria	**Exclusion criteria**
Knowledge	-Has doctoral training. -Has a considerable number of relevant academic publications such as journal articles and book chapters. - Has knowledge of e-learning, digital ethics, and/ or academic integrity together with understanding of EMI assessment procedures, and the possible AI threats.	-Does not have doctoral training -Has not published at least 5 journal articles or book chapters on relevant topics. -Does not have knowledge of e-learning -learning, digital ethics, and/ or academic integrity together with understanding of EMI assessment procedures, and the possible AI threats.
Experience	-Has a six-year period of research and university teaching. -Has at least 5 years' experience in EMI.	- Does not have a six-year period of research and university teaching. -Does not have extensive at least 5 years' experience in EMI.
Pedagogical Responsibility	-Holds a university position of pedagogical responsibility. -Has previously contributed to the design and implementation of EMI assessment procedures.	-Does not hold a university position of pedagogical responsibility. - Has not previously contributed to the design and implementation of EMI assessment procedures.

The inclusion and exclusion criteria were checked by means of an initial suitability questionnaire which potential participants were asked to complete by email. This brief survey specifically addressed each of the criteria as stipulated in Table 1 above. Candidates who met all the inclusion criteria and did not fall under any of the exclusion criteria were deemed to have the necessary expertise and were recruited as panel members for this study. Strict adherence to these pre-determined criteria allowed for objective assessment of experts.

Although the expected sample size in studies which take similar methodological approaches may include up to 300 participants (Thangaratinam & Redman, 2005), Akins et al. (2005) offer support for a much-reduced number of panellists, which can likewise achieve stability of results. The creation of the expert panel took place in accordance with the inclusion and exclusion criteria stipulated previously in Table 1. A total of 59 potential experts were identified and were sent directly to these individuals by the authors. In this email, the authors provided potential participants with a participant information sheet which stipulated the main research objectives, outlined the research design architecture, highlighted the potential benefits of participating. From the people contacted, a total of 26 people agreed to participate. Although in some instances no response was given, there were 12 potential participants who declined due to limited availability.

The resulting expert panel was composed of members from Australia, Greece, Morocco, South Africa, Spain, and the United Kingdom with a gender distribution of 14 females and 12 males. All of the 26 participants taught in EMI HE settings and had an average age of 47. The innovative research design here affords greater informant participation accounting in the initial asynchronous stage for respondents unable to contribute to the synchronous live element of the process. Hence, the asynchronous stage 1 sample is empirically greater (n=26) than that of the synchronous stage 2 (n=20).

Instruments

The two core instruments used in the study were questionnaires and focus group interviews. The initial asynchronous questionnaire contained a total of 12 items related to the focus of the study and comprised closed and open-ended questions, whilst the other questionnaires contained 33 and 11 items, respectively. The numerical variation in the subsequent questionnaire owes to the need to validate and refine key themes originating from the previous rounds.

In all three cases, the items were developed by the research team based on an extensive review of relevant literature and theory to identify key concepts and variables to measure. The questions went through an iterative process of drafting, expert review, cognitive interview pretesting, and revision over the course of two months. This process helped ensure the questions adequately measured the intended constructs and were interpreted consistently by participants. A pilot study was conducted with a sample of 10 participants from the target population. Quantitative analysis methods including exploratory factor analysis were used to evaluate the psychometric properties of the questionnaire items and identify any redundant or poorly performing questions to remove (n=2). Finally, the instrument was submitted to two independent experts for review. This validation process helped refine the final questionnaire used in the study.

Additionally, synchronous focus group sessions were conducted intermittently using Microsoft Teams. There was a total of 2 focus group sessions held during the course of the study. The focus group protocol was designed to be semi-structured, with 5 key questions developed to allow participants to reflect on and review opinions on the study's focus. The focus group questions were informed by previous responses in the earlier stages of the study and underwent expert review by three experienced qualitative researchers to evaluate clarity, relevance, and likely effectiveness in generating discussion. The questions were also revised based on the feedback prior to the sessions.

Data Collection

This study utilised a mixed methods approach to data collection, including both quantitative and qualitative techniques carried out over multiple iterative rounds. Quantitative data were gathered through closed-ended questionnaire items administered via three rounds of Delphi questionnaires. This allowed for gathering statistical data reflecting experts' consensus and dissent on key topics. In addition, qualitative data were collected through open-ended questions included in each Delphi questionnaire round, as well as through two interjacent synchronous focus group sessions conducted online between consecutive rounds. This multi-stage process enabled experts to engage in increasing depth and reflectivity on the topics, moving from broader commentary to more focused insights.

Each stage of data collection took responses from the previous round and allowed experts to consider, elaborate and refine these in the next stage. Items which did not achieve substantial agreement were disregarded, as is detailed subsequently in the Data Analysis section. The process concluded when expert consensus was reached on all the items, and no further refinements or elaborations were suggested by the panellists. To facilitate this, aligning with relevant methodological literature (Dörnyei and Taguchi, 2009), the questionnaires intentionally progressed from general closed and open-ended items in Round I towards more specific and targeted quantitative and qualitative questions in the latter rounds. This general-to-specific approach provided structure to funnel experts' opinions while allowing flexibility for elaboration through the open-ended and focus group data.

The combination of asynchronous and synchronous data collection, increasingly focused questioning, and multiple opportunities for reflection were designed to foster rich, multi-faceted insights from experts through both numerical ratings and descriptive commentaries. The mixed methods Delphi process aimed to produce comprehensive, trustworthy data to address the complex research phenomena at hand.

Data Analysis

Firstly, the numerical definition of expert consensus in the quantitative data was sought by the researchers. Throughout, consistent with methodological literature (Diamond et al., 2014) and the limited sample size (von der Gracht, 2012), Cohen's kappa coefficient (ϰ) was employed to establish the strength of agreement. The established benchmark for consensus for the expert panellists for each item was equal to or above a kappa (ϰ) value of 0.75 (Detorri & Norvell, 2020) and items with a kappa coefficient (ϰ) of < 0.74 were consequently not carried over into the subsequent phase of the research procedure. To this end, the information provided in Table 2 below was used for interpretation:

Table 2
Cohen's kappa Coefficient Interpretation for Strength of Agreement

Cohen's kappa coefficient (ϰ)	Strength of agreement
< 0.00	Poor agreement
0.00 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

From this, the quantitative data collected in this study were then analysed using SPSS Statistics software (version 29.0) employing descriptive statistics to measure the level of consensus of the expert panellists. Cohen's kappa efficient was used to measure inter-rater reliability as is detailed previously. The rationale for this was that two coders were involved in the data analysis procedure of the study and as per documented literature, Cohen's kappa is particularly suitable when there are two coders (Hsu & Sandford, 2007).

On the other hand, the qualitative data were analysed using thematic analysis (Braun et al., 2019). Initial open coding was conducted by two independent researchers to extract salient themes from the data. The two researchers individually reviewed the transcripts, identified preliminary themes, and assigned codes to relevant passages of text. After initial coding, the two researchers met to compare their identified themes and codes. Any discrepancies were discussed until consensus was reached to align coding decisions.

To evaluate inter-coder reliability, a Cohen's kappa coefficient was calculated between the two coders' theme assignments. The preliminary themes were further refined and organised into major categories through an inductive, data-driven process. The researchers worked collaboratively to group related themes into higher-order categories based on underlying relationships and patterns observed in the data. Coded data were compared against the quantitative results through data triangulation (Rothbauer, 2008), allowing convergence and corroboration between the quantitative statistics and qualitative themes to provide a comprehensive understanding of the research questions. Together, the quantitative and qualitative analyses provided complementary insights into participants' experiences and perspectives related to the study aims. The integration of both forms of data and analysis allowed for a more complete and nuanced interpretation of results than either method alone.

Results

The results reflect the process of idea building, refinement, and expert consensus consolidation, which was undertaken to address RQ1 and RQ2, and ultimately create the GenAI and EMI Assessment Problem-Solution Matrix as is presented in the following section.

Delphi Round I Questionnaire

This entry point questionnaire yielded validation for the raison d'être of the present study by conferring expert consensus on the potential of GenAI usage in EMI pedagogical contexts both as a threat to current EMI assessment praxis and as an opportunity to rethink assessment altogether (Atlas, 2023), achieving almost perfect agreement ($\varkappa$= 1.00) on both counts. However, divergence, or no expert consensus, was found to converge on current international HE provision to deal with potential present and future AI-related threats or to reenvisage an EMI assessment model fit for purpose, respectively yielding slight agreement ($\varkappa$= 0.125) and fair agreement ($\varkappa$= 0.25).

Furthermore, in response to RQ1, informants answered open-ended questions to identify potential threats to EMI assessment integrity. Participant responses yielded a total of 83 items pertaining to present (n=59) and future (n=24) issues and threats to academic integrity in EMI assessment. The most salient and frequent examples included 'AI-assisted online exam cheating', 'plagiarism', 'AI-powered ghost-writing in formative assessments', and 'difficulty to distinguish student writing from that of GenAI tools'. Qualitative data gathered from these open-ended questions were then subjected to thematic analysis and were organised into 32 key themes from the 74 codes identified.

The themes identified included: 'widespread GenAI use by students for generating text', 'students using GenAI to create work with little effort', 'GenAI increasing incidents of academic misconduct', 'GenAI making plagiarism and cheating easier', 'issues with GenAI not well known by institutions', 'lack of policies about GenAI use in assessments', 'institutions unprepared to address GenAI cheating', 'GenAI use impossible to fully prevent in assessments', 'limitations in designing out GenAI misconduct', 'GenAI an attractive shortcut for busy EMI students', 'EMI students drawn to GenAI's language capabilities', 'entry requirements misaligned with EMI course demands', 'EMI courses linguistically overwhelming for some students', 'open-book exams particularly risky due to GenAI text generation', 'dissertations vulnerable to GenAI plagiarism', 'non-proctored assessments susceptible to GenAI cheating', 'institutions quickly banning GenAI use in assessments', 'severe penalties might be instituted for GenAI cheating', 'assessment policies shifted in reaction to GenAI risks', 'other AI tools also posing academic integrity risks', 'additional GenAI capabilities requiring consideration', 'modular courses reducing understanding of student writing', 'fragmented assessments in modular courses', 'peer interaction lost in modular EMI programmes', 'isolated modules enabling undetected GenAI use', 'authenticity and authorship issues in modular assessments', 'challenges verifying student work in modular courses', 'integrated course design as deterrent to GenAI cheating', 'longitudinal analysis of student writing in integrated courses', 'comprehensive GenAI misconduct strategy needed', 'holistic multi-stakeholder approach required', and 'nuanced GenAI policies for balanced mitigation. These themes were then the subject of discussion in Focus Group I.

Focus Group Session I

Again, addressing RQ1, in the synchronous discussion with a reduced number of available expert panellists (n=20) the panellists reviewed the feedback from the first stage, and 29 codes were identified by the researchers in total and from these 9 key themes were defined. These items were triangulated with the data collected previously. There were a number of diverse concerns articulated by participants:

> *"In my experience, there's a real lack of understanding in institutions about how to address students using AI for academic misconduct both generally and even more so for us in EMI. I just don't agree with you I'm afraid. Policies haven't kept pace with these emerging technologies." [FG.1.A46]*

> *"No matter how carefully assessments are designed, students motivated to misuse AI will find ways to do it. We can't completely design out the risk." [FG.1.F108]*

> *"With the time pressures on EMI students, I can understand why some see AI text generators as a shortcut to meet assignment requirements, even if unethical." [FG.1.G17]*

> *"As we all know by now, there's often a mismatch between English requirements for admission versus the actual language skills needed for assignments. This disparity creates challenges." [FG.1.M129]*

> *"In my view, take-home exams and dissertations without proctoring are highly likely to be vulnerable. I mean you could get ChatGPT to do the whole thing if you really wanted to." [FG.1.R52]*

> *"I see these comments on cracking down, but I worry institutions will overcorrect with harsh policy changes that disadvantage honest students, in an effort to combat AI misconduct." [FG.1.R114]*

> *"With modular systems, lecturers have less exposure to individual students' writing. In EMI the development of their writing is half the battle of course. And now this makes AI misconduct harder to detect." [FG.1.B81]*

From their contributions during the session, the themes which were identified included 'AI-assisted academic misconduct incursion', 'lack of institutional awareness and readiness to deal with the issue', 'AI-assisted academic malpractice cannot be designed out of assessment', 'tempting shortcut for busy EMI students', 'disparity between language entry requirements and course reality', 'open-book exams, dissertations and non-human proctored assessments particularly susceptible', 'drastic and disproportionate institutional assessment policy shifts', 'further AI-powered tools need addressing', and 'modular HE not conducive to development of awareness of student writing characteristics'. Therefore, the triangulation of the data at this point formed the refined foundations on which the challenges posed by AI text generators for EMI assessment for the proposed problem-solution matrix were determined.

Delphi Round II Questionnaire

Informed by the results from the previous stages, the researchers strived to consolidate expert consensus on the GenAI-related challenges for EMI assessment in this stage. They aimed to facilitate the generation of ideas amongst the expert panellists related to the conceptualisation of EMI assessment recommendations to address such threats to academic integrity.

The resulting quantitative data yielded showed almost perfect agreement (ϰ= 1.00), thereby expert consensus was reached on the following issues: 'AI-assisted academic misconduct incursion', 'lack of institutional awareness and readiness to deal with the issue', 'disparity between language entry requirements and course reality', 'drastic and disproportionate institutional assessment policy shifts', 'further AI-powered tools need addressing', and 'modular HE not conducive to development of awareness of student writing characteristics'.

Further consensus with substantial agreement (ϰ= 0.80) was established regarding 'tempting shortcut for busy EMI students' and 'open-book exams, dissertations and non-human proctored assessments particularly susceptible'; and finally, however, there was expert panellist divergence with only fair agreement (ϰ= 0.40) on the matter of 'AI-assisted academic malpractice cannot be designed out of assessment'. Confirming earlier misgivings raised in the Introduction and in accordance with RQ1, this refinement of ideas led to the establishment of 8 definitive challenges posed by AI for EMI assessment to be included in the proposed problem-solution matrix as are detailed in Table 3 below:

Table 3
Delphi Round II Results

GenAI-related EMI Assessment Issue of Concern	Strength of Agreement
AI-assisted academic misconduct incursion	Almost perfect agreement (ϰ= 1.00)
Lack of institutional awareness and readiness for unauthorised AI tool usage detection	Almost perfect agreement (ϰ= 1.00)
Disparity between required student linguistic proficiency and academic skillset and university language entry requirements.	Almost perfect agreement (ϰ= 1.00)
Drastic institutional policy shift considering agitation in current landscape, e.g., a return to final exam assessment diets.	Almost perfect agreement (ϰ= 1.00)
Range of other AI-powered tools which may exploit loopholes in current institutional detection mechanisms, such as for paraphrasing and interlanguage translation.	Almost perfect agreement (ϰ= 1.00)
Modular nature of HE at present is not conducive to the development of awareness of individual student writing characteristics along the student journey.	Almost perfect agreement (ϰ= 1.00)
Tempting shortcut for busy students in EMI settings	Substantial agreement (ϰ= 0.80)
Open-book style exams, dissertations and other non-human proctored assessment components may be particularly susceptible to AI-assisted academic misconduct.	Substantial agreement (ϰ= 0.80)

Moreover, the qualitative data gathered from the open-ended questions pertaining to the conceptualisation of possible recommendations to address AI-related threats to EMI assessment (RQ2), were then subjected to thematic analysis. A total of 38 codes were defined and the researchers classified these into 11 key themes. These included: 'institutional policy development', 'raising awareness of all key stakeholders', 'enhancement of digital tools for detection', 'staff-student ratio re-evaluation', 'greater EAP programme and curriculum integration', 'greater clarity of academic misconduct consequences', 'greater institutional investment in EAP pre-sessional and in-sessional programmes', 'development of authentic assessment alternatives based on higher order thinking skills (HOTS)', 'English language assessment criterion in assessments', 'creation of working groups to monitor', and 'HE instruction and assessment transformation'.

Focus Group Session II

In keeping with RQ2, the resulting key themes were then subject to discussion in Focus Group Session II. There was some notable disagreement, particularly regarding the terminology used in the key theme of 'institutional policy development' and its validity as a potential recommendation as part of the problem-solution matrix. To that end, it is comprehensible that of the 20 codes identified and the 8 key themes which they were organised into, 4 contributed to the identification of a new theme, i.e., 'institutional policy and guidance documentation development'. The following extracts illustrate in part some of the discussion on this point:

> *"I'm concerned using the term 'policy' may imply top-down edicts. Guidance developed collaboratively with faculty may be received better." [FG.2.C26]*

> *"In my institution, policy comes from the administration while guidelines originate in schools/departments. We should consider terminology that reflects this." [FG.2.P27]*

> *"But is 'institutional policy development' the right framing though? To me, that connotes bureaucratic mandates. I prefer 'guidance documentation' as it's more flexible." [FG.2.C28]*

> *"Well realistically each university will have its inner workings which I feel fall beyond the scope here. How about we include both terms?" [FG.2.H29]*

In addition, there were other comments of note made pertaining to potential recommendations, for instance:

> *"Raising awareness across all university groups - students, faculty, leadership is crucial for tackling this challenge." [FG.2.O62]*

> *"We need better digital tools to detect AI-generated text if we want to curb misconduct." [FG.2.J89]*

> *"With rising enrolments and reliance on AI, we must rethink appropriate staff-student ratios." [FG.2.B111]*

> *"Stronger integration of English language curriculum in academic programs is key to setting students up for success with assignments." [FG.2.C148]*

> *"Students need to understand clearly the serious consequences of using AI writing tools unethically." [FG.2.165]*

> *"Institutions should invest more in English language support throughout students' studies, particularly if they are seriously committed to EMI going forward." [FG.2.G177]*

> *"We can't talk about future-proofing anything with the way things are going in AI. We should develop assessments focused on higher-order skills, not easily replicated by AI and bin the essay once and for all." [FG.2.L199]*

Ultimately, the key themes identified at this point were: 'institutional policy and guidance documentation development', 'raising awareness of all key stakeholders', 'enhancement of digital tools for detection', 'staff-student ratio re-evaluation', 'greater EAP programme and curriculum integration', 'greater clarity of academic misconduct consequences', 'greater institutional investment in EAP pre-sessional and in-sessional programmes', 'development of authentic assessment alternatives based on higher order thinking skills (HOTS)', 'English language assessment criterion in assessments', 'creation of working groups to monitor', and 'HE instruction and assessment transformation'.

Delphi Round III Questionnaire

Again addressing RQ2, the quantitative and qualitative data gathered previously were triangulated and were taken into account in the creation of the subsequent questionnaire, which aimed to gather data on potential recommendations to deal with potential AI-assisted academic misconduct in EMI pedagogical settings. Surprisingly, expert consensus was achieved with almost perfect agreement ($\varkappa$= 1.00) on 8 of the recommendations and with substantial agreement ($\varkappa$= 0.80) on the remaining 3.

Despite this, the thematic analysis of the resulting qualitative data highlighted 4 codes, which were organised into 1 key theme: 'concerns around HE readiness and provision to enact necessary change'- this point is subsequently explored in greater depth in the following section. As per RQ2 and addressing the previously highlighted gaps in EMI assessment extant literature and praxis, the confirmatory expert consensus established at this point represented the conclusion of the process and conferred definitive status to the EMI assessment recommendations to be used in the proposed problem-solution matrix. This information is detailed in Table 4 below.

Table 4
Delphi Round III Results

Proposed EMI Assessment Solution	Strength of Agreement
Raise awareness amongst key stakeholders on responsible use of AI tools in academia.	Almost perfect agreement ($\varkappa$= 1.00)
Exploration of further digital tools to enhance AI-related plagiarism detection repertoire.	Almost perfect agreement ($\varkappa$= 1.00)
Curriculum integration on EAP and content formative programmes on authorised AI tool praxis in academia.	Almost perfect agreement ($\varkappa$= 1.00)
Greater dissemination on consequences of academic misconduct if detected.	Almost perfect agreement ($\varkappa$= 1.00)
Further investment and development of English for General and Specific Academic Purposes in-sessional and pre-sessional programmes.	Almost perfect agreement ($\varkappa$= 1.00)
The inclusion of English language usage assessment in assessment criteria and collaboration with EAP specialists to this end.	Almost perfect agreement ($\varkappa$= 1.00)
Development of a wider range of authentic assessment tasks in-house which enable students to employ higher order thinking skills in context.	Almost perfect agreement ($\varkappa$= 1.00)
Pursuit of institutional shift towards alternative modes of EMI tuition and assessment praxis.	Almost perfect agreement ($\varkappa$= 1.00)
Specific AI-related institutional policy and guidance documentation development.	Substantial agreement ($\varkappa$= 0.80)
Re-evaluation of student-staff ratios for greater time for academic and assessment literacies dialogues to occur.	Substantial agreement ($\varkappa$= 0.80)
Creation of faculty and multidisciplinary AI-assisted academic misconduct institutional working groups.	Substantial agreement ($\varkappa$= 0.80)

Discussions

The impetus of this study was to ascertain expert consensus on potential threats from AI-powered text generation tools and on possible recommendations to counteract these in EMI HE Assessment. Thus, RQ1 and RQ2 were formulated with the purpose of crafting a problem-solution matrix for the sector to this end. The subsequent research design enabled the fulfilment of this and yielded a wealth of quantitative and qualitative data, which was subsequently refined and used to populate the AI and EMI Problem-Solution Matrix. The findings here respond to the EMI research agenda highlighted previously in the Literature Review section (Sah, 2022) The results are practical in nature but also key into underlying theoretical assumptions and principles without specifically attending to the particularities of any given GenAI application in alignment with the stance taken by Passey (2019) outlined previously in the Literature Review section.

Throughout this process, limited expert divergence was perceived, and interestingly, there was stability of expert consensus that was conferred on both the issues and recommendations conceptualised with the final iteration of items achieving either substantial agreement (ϰ= 0.80) or almost perfect agreement (ϰ= 1.00). Drawing on these novel findings, the key stakeholders in the sector may now embrace the recommendations and/or carve a bespoke route to deal with the issues identified.

Contributions to Extant Literature

The GenAI and EMI Assessment Problem-Solution Matrix constitutes a novel practical contribution to the field for key stakeholder use, which goes some way to address the numerous gaps in extant literature identified in the Literature Review section. It is, however, by no means a definitive outcome given the ever-changing and fast-developing nature of AI-powered tools and the emerging challenges they may continue to present, but rather, the authors share this with the wider academic community in the hope of making a contribution to extant literature which will go on to generate much-needed scholarly debate and further consensus conformation on the challenges presented and how to deal with them. Moreover, it is hoped that the outcome here may be of some use in the development of clear pedagogical support, which, as noted earlier in the Introduction, is often found to be lacking in EMI HE settings (Farrell, 2019).

In the Literature Review section, it was also emphasised that GenAI text generation tools, as is the case with the rise of other digital technologies, create new challenges for maintaining academic integrity in higher educational institutions (Crawford et al., 2023). The results here seemingly corroborate this stance, showing that GenAI tools may facilitate academic misconduct, and students may use them as shortcuts. They also key into the specific challenges of ‘problematic EMI assessments (Dearden, 2014) and illustrate further challenges in addition to those conceptualised by Inbar-Lourie (2022) and Li and Wu (2018), such as assessing students in a language they are still learning, meaning that assessment results may not accurately measure students' acquisition of subject knowledge due to potential interference from English proficiency (Robinson, 2010; van der Walt & Kidd, 2013). The results specifically address this point by recommending greater provision for EGAP/ESAP support. Although this may be helpful to students in practice, it does not resolve the ongoing assessment dilemma in EMI contexts. Moreover, a wider range of further challenges that threaten EMI assessment integrity have now been conceptualised as detailed in the matrix, illustrating that the emergence of GenAI writing tools have added to the overall complexity of EMI assessment and perhaps made this even more problematic than previously thought. There is alignment between findings and the calls in scholarship for the need for guidelines and policies on the responsible use of AI in education (Chan, 2023; Lodge et al., 2023) to address GenAI-related issues for EMI assessment integrity to contribute towards the mitigation of risks.

The issues and recommendations identified in this study hold relevance for the particularities of EMI HE, for example, the specific need for greater EAP pre-sessional and in-sessional provision and the shift towards EMI assessment alternatives that move away from the traditional essay, such as multimodal assessments or the concept of ungrading. There is application too beyond the specific setting examined. For instance, the expert panel highlighted challenges like insufficient instructor preparation and lack of plagiarism detection tools as widespread concerns. Their suggested solutions, like teacher training and updated academic integrity policies, could be applicable across many HE institutions grappling with emerging AI text generator technologies. Furthermore, while this study focused on the HE level, the extensive skills needed for academic English and writing are central in secondary and vocational education contexts as well. Therefore, the universal nature of the GenAI threats and EMI assessment recommendations conceptualised in this study provides initial evidence that they may generalise to other international educational settings using EMI, from secondary schools to vocational training programmes. Additional research would be valuable to further investigate the applicability of these findings for enhancing academic integrity practices and EMI assessment in diverse educational contexts.

In addition, the fundamental assessment principles put forward by the experts may point to the applicability of certain issues and recommendations beyond EMI assessment contexts. That is to say, the suggestions to create cross-disciplinary working groups, implement more authentic alternatives to traditional assessments, reduce educator workloads, orientate student GenAI tool usage, and examine possibilities beyond the modular structure of HE all relate to broader assessment challenges. Though this study focused specifically on EMI programme assessments, these underlying assessment principles transcend the EMI context and the applicability of both the challenges and the recommendations in the matrix may impact university policy for both EMI and non-EMI HE contexts. The consensus reached provides initial evidence that the issues and recommendations could be of use to enhance assessment practices in non-EMI HE contexts as well. Further research would be valuable to explore the generalisation of these foundational assessment elements to other academic disciplines and contexts facing similar assessment needs and issues in an age of advancing GenAI technologies. .

Addressing the present understandably barren scholarly landscape, owing to the novelty of the GenAI phenomenon, as highlighted previously in the Literature Review section, this timely study has made a novel contribution to address the gaps identified. Moreover, the novel methodological approach taken has limited precedent in the field. The successful accomplishment of the researchers' aims which overcame the aforementioned shortcomings articulated by Chan (2022) in the Methods section may therefore go some way to validate the adaptations stipulated previously. These presumably beneficial improvements enhance the case for broader application of the approach in the fields of EMI, Assessment and Evaluation, and HE investigation.

The Role of EMI Assessment in Higher Education

Drawing on earlier references to this point in the opening lines of this article, the role of EMI assessment in HE was scrutinised on the matter of tensions between testing and assessment. These, in turn, were echoed on several occasions by informants and are arguably manifested in the recommendations yielded. In particular, these can be found in the recommendations pertaining to authentic assessment design and other formative alternative exploration in the problem-solution matrix that has been developed in this study.

This social constructivist approach sits well with the numerous EMI assessment recommendations which have been conceptualised in the AI and EMI Assessment Problem-Solution Matrix in this study. These consider the possibility for meaningful dialogue among stakeholders (Sadler, 2010), greater precision in feedback (Carless, 2015), and a deeper approach to learning (Entwistle, 2018). In short, recommendations conceptualised represent serious challenges to the hegemonic commodification and subsequent massification of HE and its implications for social justice (Milligan & Tikly, 2018), that could be used to mitigate the potential amplification of such issues through GenAI privileged students to artificially boost their academic performance (Mortenson, 2021). Proactive policies and alternate assessment approaches as mentioned in the matrix, are needed to promote equitable EMI assessment in the age of GenAI. This stance was also articulated by 2 participants, and the results, as well as the subsequent debate. The findings here would suggest that there are many paths to be explored to solve this internally.

The Delphi method used in this study shows promise for garnering support for the development of alternative assessment provision, including learning-orientated and formative assessments presently found to be lacking in practice as cited earlier in the Literature Review section (Inbar-Lourie, 2022; Li & Wu, 2018). Convening busy experts in multiple rounds of anonymous discussion allowed for interactive, constructive consensus building. As highlighted in the Materials and Methods section, the Delphi technique channels "wisdom-of-crowds" insights by synthesizing perspectives from a diverse panel of specialists (Surowiecki, 2005). Though time-intensive, this structured group communication process produced clear guidelines and priorities for enhancing EMI assessment in light of AI text generators. The experts specifically recommended authentic integrated tasks, project-based assessment, viva voce exams, and assessing both process and product. Such alternative assessments could help promote academic integrity and English proficiency development. This initial Delphi study demonstrates that directly eliciting recommendations from experienced EMI educators and applied linguists can yield tangible, feasible improvements for evaluation practices. Further research is warranted to implement and evaluate the proposed alternative assessments in educational contexts. However, this study indicates that the Delphi method's structured, collaborative approach can generate targeted assessment solutions. .

Further support may be drawn from the findings here. In keeping with this concept of assessment development in community, Rolfe (2013) proposes resisting managerialism and the accompanying predominance of the associated definition of 'productivity' by banding together to protect academically valued practices. This viewpoint is consistent with the participants' worries about present HE provision's ability to handle the difficulties posed by AI-powered digital technologies in a timely and efficient manner. Moreover, the findings of this study are consistent with the central notion of 'living in the ruins of HE', which is based on Readings' (1997) posthumously released work. As a result, bringing contemplation and open discussion back to the forefront of academic work runs against to the corporate conception of scholarly output. In other words, taking the time to contemplate and discuss ideas with others may be an effective way to plant seeds of change.

In sum, a HE institution can and should be far greater than vested economic interests; it is its people, their points of view, and thoughtful exchanges irrespective of the language used as a vehicle for communication. Only once this becomes universally accepted, and all sectors are able to unite in both thought and action, can substantive alterations in the role of EMI assessment, and the wider HE sector all-round, come into being in the shadow of GenAI.

Limitations

The advantages of the chosen methodological approach were highlighted earlier, but there are some limitations that could affect the results' credibility by reducing their validity and reliability. An expert panel was formed as a result of the inclusion and exclusion criteria used in the selection and recruitment processes undertaken, which were described previously. Although the panel members who participated in this study satisfied these requirements, as mentioned previously in the Literature Review section, Fink-Hafner et al. (2019) emphasise that the outcomes might alter if the exercise were performed with different panel members. Although care has been taken to remind the reader that the project is exploratory, it is advised that more research be done to verify the findings to remedy this.

The utilisation of focus group sessions as a component of the exploratory international modified Delphi technique is another restriction. The bandwagon effect, susceptibility to manipulation, and hesitation to change one's opinion in front of others are a few well-known concerns at play here (Greenbaum, 1998). These problems are particularly pressing in the current study since it relies heavily on opinion expression and change to build and negotiate the body of information that serves as the foundation for expert consensus. At the final stage of the synchronous session, where participants completed the Delphi Round 3 Questionnaire, a potential danger of this may be emphasized. In other words, to avoid future rounds of discussion and in reaction to the timing of this research phase, there is a risk that participants might have just expressed conformity to the final iteration of the matrix.

Future Lines of Investigation

The researchers are confident that the results of this study have taken a step towards bridging the gaps previously identified and hope that findings will spark further research endeavours amongst the global scholarly community. However, there is a wide range of further areas of investigation to be explored and suggestions made here by no means are intended to be exhaustive.

Amongst these, key lines to be highlighted include the further validation and refinement of results not only to address the methodological limitations stipulated previously but also to ensure that the GenAI and EMI Assessment Problem-Solution Matrix is kept updated in line with the latest developments of the field. These should ideally key into the tenets of the informing assessment and pedagogical principles behind these as opposed to specifically exemplifying with the latest tools *du jour*. Given the advances here in terms of equitable EMI assessment, it is clear that further investigation ought to be made to ensure that the possible associated risks in practice are documented, and further bespoke solutions are developed.

Moreover, the development of an institutional AI-related EMI assessment policy together with supplementary guidance documentation, as recommended by the expert panel is an additional avenue of exploration given the extremely limited information available in the sector, as was raised previously (e.g., UCL, 2023). In addition, further investigation might also address the implications for teacher education programmes and EAP-specific professional development programmes such as the BALEAP TEAP Individual Accreditation Scheme (BALEAP, 2022). Evidently, further work into student use and delving into the reasons behind this would also be helpful all round, together with exploration of how the use of GenAI-powered tools might enrich EMI classroom best practices.

The GenAI and EMI Assessment Problem-Solution Matrix

In fulfilment of the principal research objective, the culmination of the investigative efforts founded on expert knowledge building, idea refinement, and consensus consolidation to address the discursive gap identified and problematised in the RQs is presented as the GenAI and EMI Assessment Problem-Solution Matrix as per Table 5 below:

Table 5
The GenAI and EMI Assessment Problem-Solution Matrix

AI-related EMI Assessment Issues	Proposed Recommendation
AI text generation software may facilitate academic malpractice incursion in EMI contexts.	-Development of a comprehensive institutional AI assessment policy and accompanying guidance which consider the multiple pedagogical realities of HE today and informed by input from all key stakeholders.
Current institutional provision, such as lecturer awareness of tool availability, assessment literacy and student use, and dedicated marking time, together with digital tools e.g. Turnitin, may not be sufficient to detect student AI tool use in assessment.	-Raise awareness amongst key stakeholders in the HE community and develop a training programme, such as a series of seminars, on the responsible use of AI tools and how they may be used to enhance learning. -Exploration of further digital tools and enhancement of current provision to develop more effective digital detection of AI-related plagiarism. Liaison with and lobbying leaders in the digital community to this end. -Re-evaluation of student-staff ratios to enable more time for conscious academic and assessment literacies dialogue.
AI text generation applications represent a tempting assessment shortcut particularly for busy students in demanding EMI pedagogical contexts.	-Curriculum integration on EAP and content formative programmes to educate students on how such tools should be used. -Further clarity in institutional guidance on this matter in relation to academic misconduct and its consequences when detected.
Students who meet university language entry requirements may in fact not have the linguistic proficiency and academic skill set to meet expected assessment standards and requirements and deem applications such as ChatGPT a plausible alternative.	-Greater investment and provision for English General and Specific for Academic Purposes (EGAP and ESAP) in-sessional and pre-sessional programmes of tuition in which the responsible use of AI could be addressed as part of formative academic and assessment literacy preparation and support for proficient and L2 users of English.
Risk of institutional and sector-wide malaise and agitation due to AI tools and the threat to academic integrity, possibly provoking drastic shifts in assessment policy, e.g., a return to final exam assessment diets.	-Development of authentic assessment tasks in-house which enable students to employ higher order thinking skills in practice in line with greater alignment with QA Subject Benchmarking requirements and professional demands of their chosen field.
Open-book exam questions, dissertations and other means of assessment not carried out under invigilated exam conditions may be susceptible to the use of ever-increasing AI text generation tools.	-Inclusion of English language assessment in assessment criteria, with input from EAP specialists where possible both pre-, during, and post-assessment.
Availability of other machine translation software tools of concern such as Quillbot for paraphrasing and DeepL for automatic interlanguage translation, which may exploit loopholes in current institutional academic integrity preservation detection mechanisms.	-Creation of faculty and multidisciplinary institutional working groups to closely monitor and act upon AI-related incidences of malpractice.
Seemingly lack of pedagogical continuity in the modular nature of HE and subsequent student anonymity in the student learning journey at present. This does not allow for lecturers to develop an awareness of individual student writing characteristics, thus augmenting the challenge for AI text generation detection.	-Institutional shift towards innovative multi-module learning outcome-focused assessment praxis, including, but not limited to, integrated assessments, 'ungrading', multiple phases of assessment e.g., formative marking of drafts, group and project-based assessment, portfolios, and multimodal assessments.

Conclusion

This exploratory study set out to delve into the possible GenAI-related threats for EMI assessment and proposals to remedy these in the context of HE. The authors were met with an unprolific scholarly landscape in which a lack of expert consensus in extant literature and praxis was to be found, seemingly by virtue of the fast-paced and emergent nature of the matter at hand. To that end, the creation of a problem-solution matrix for key stakeholders was proposed and crafted thanks to the novel adaptation of the established but seldom employed Delphi method. The international modified Delphi study yielded qualitative and quantitative data from expert panellists and the culmination of these constituted the creation of the AI and EMI Assessment Problem-Solution Matrix.

The novel scholarly exploration undertaken in this study has arguably made a theoretical and practical contribution to the field with potential further application in other didactic settings. This tool intends to be used as a means of social justice empowerment which contributes to creating a more level playing field for EMI students founded on authentic and meaningful assessment procedures in which GenAI tools are used in an equitable way and thus promote a more inclusive learning environment for all. Limitations have been duly acknowledged together with suggested further lines of investigation which are by no means exhaustive. The approach taken here has permitted the authors to reflect on and question the wider role of assessment within the massified and commodified model of HE today and identify further systemic and paradigmatic limitations and challenges afoot for the international academic community.

In sum, as has been evidenced here, the challenges presented by AI-powered tools may not actually signify end game for EMI and, arguably, non-EMI, assessment academic integrity, but rather, it is only when educators of all stripes step out of the shadows and come together in community to openly discuss and reflect on the necessary underpinning systemic pedagogical changes, that we may in fact usher in the dawn of a new didactically fruitful era.

Funding Acknowledgment and Ethics Approval

This research has been carried out as part of the Project of Analysis and Development for the Optimization of Assessment and Regulation of Generative Artificial Intelligence in Humanities (PANDORA), with project reference number PP-2023-22, financed by Universidad Internacional de La Rioja (UNIR), Spain.

The study was conducted having been granted ethics approval from Universidad Internacional de La Rioja (UNIR) with ethics approval code PI006/2023.

References

Abd-Elaal, E.-S., Gamage, S. H. P. W., & Mills, J. E. (2022). Assisting academics to identify computer-generated writing. *European Journal of Engineering Education*, *47*(5), 1-21. https://doi.org/10.1080/03043797.2022.2046709

Airey, J. (2012). "I don't teach language". *AILA Review*, *25*, 64-79. https://doi.org/10.1075/aila.25.05air

Akins, R. B., Tolson, H., & Cole, B. R. (2005). Stability of response characteristics of a Delphi panel: Application of bootstrap data expansion. BMC Medical Research Methodology, 5, 1-12. https://doi.org/10.1186/1471-2288-5-37

Atlas, S. (2023). *ChatGPT for Education and Professional Development: A guide to conversational AI.* The University of Rhode Island. https://digitalcommons.uri.edu/cba_facpubs/548/

BALEAP (2022). *BALEAP TEAP Individual Accreditation Scheme.* https://bit.ly/3lY1deK

Bishop, L. (in press). A Computer Wrote this Paper: What ChatGPT Means for Education, Research, and Writing. https://dx.foi.org/10.21.39/ssrn.4338981

Braun, V., Clarke, V., Hayfield, N. & Terry, G. (2019). Thematic analysis. In P. Liamputtong(Ed.), *Handbook of Research Methods in Social Sciences* (pp. 843-860). Springer. https://doi.org/10.1007/978-981-10-5251-4_103

Cabaleiro-Cerviño, G., & Vera, C. (2020). The Impact of Educational Technologies in Higher Education. *GIST Education and Learning Research Journal, 20*, 155-169. https://bit.ly/3kpZcYm

Carless, D. (2015). *Excellence in university assessment: Learning from award-winning practice.* Routledge. https://doi.org/10.4324/9781315740621

Carless, D., Bridges, S. M., Chan, C. K. Y., & Glofcheski, R. (Eds.) (2017). *Scaling up assessment for learning in Higher Education.* Springer. https://doi.org/10.1007/978-981-10-3045-1

Chan, C. K. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00408-3

Chan, P. (2022). An empirical study on data validation methods in Delphi and general consensus. *Data*, *18*, 1-18. https://doi.org/10.3390/data7020018

Chen, N. N., & Saulter, R. (2019). Accommodating writing tests for second language learners with disabilities. In T. Ruecker, & D. Crusan (Eds.), *The Politics of English Second Language Writing Assessment in Global Contexts* (pp. 170-185). Routledge.

Chew, E., Ding, S. L., & Rowell, G. (2015). Changing attitudes in learning and assessment: Cast-off "plagiarism detection" and cast-on self-service assessment for learning. *Innovations in Education and Teaching International, 52*, 454. https://doi.org/10.1080/14703297.2013.832633

Coyle, D., Hood, P., & Marsh, D. (2010). *Content and language integrated learning.* Cambridge University Press.

Crawford, J., Cowling, M., & Allen, K. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching and Learning Practice*, *20*(3). https://doi.org/10.53761/1.20.3.02

Cuhls, K. (2001). Foresight with Delphi Surveys in Japan. *Technology Analysis & Strategic Management*, *13*, 555-569. https://doi.org/10.1080/09537320127287

Dai, Y., Liu, A., & Cher Ping Lim. (2023). Reconceptualizing ChatGPT and generative AI as a student-driven innovation in higher education. *Procedia CIRP*, *119*, 84–90. https://doi.org/10.1016/j.procir.2023.05.002

Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. *The International Journal of Management Education*, *21*(2). https://doi.org/10.1016/j.ijme.2023.100822

Dearden, J. (2014). *English as a medium of instruction – a growing global phenomenon*. University of Oxford. https://bit.ly/3IGRp1I

Denisova-Schmidt, E. (2017). *The challenges of academic integrity in higher education: Current trends and prospects*. Boston College Center for International Higher Education. https://bit.ly/3IH2xeZ

Detorri, J. R., & Norvell, D. C. (2020). Kappa and beyond: Is there agreement? *Global Spine Journal*, *10*, 499-501. https://doi.org/10.1177/2192568220911648

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting Delphi studies. *Journal of Clinical Epidemiology*, *67*, 401-409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Doiz, A., & Lasagabaster, D. (2018). Teachers' and students' second language motivational self system in English-medium instruction: A qualitative approach. *TESOL Quarterly*, *52*(3), 657-679. https://doi.org/10.1002/tesq.452

Dörnyei, Z. & Taguchi, T. (2009). *Questionnaires in Second Language Research*. Routledge. https://doi.org/10.4324/9780203864739

Entwistle, N. (2018). *Student learning and academic understanding: A research perspective with implications for teaching*. Academic Press. https://doi.org/10.1016/b978-0-12-805359-1.00012-7

Escotet, M. Á. (2023). The optimistic future of artificial intelligence in higher education. *PROSPECTS*. https://doi.org/10.1007/s11125-023-09642-z

Farrell, T. S. (2019). Professional development through reflective practice for English-medium instruction (EMI) teachers. *International Journal of Bilingual Education and Bilingualism*, *23*(3), 277-286. https://doi.org/10.1080/13670050.2019.1612840

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating Academic Answers Generated Using ChatGPT. *Journal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.3c00087

Fink-Hafner, D., Dagen, T., Doušak, M., Novak, M., & Hafner-Fink, M. (2019). Delphi method strengths and weaknesses. *Advances in Methodology and Statistics*, *16*, 1-19. https://doi.org/10.51936/fcfm6982

Green, R. (2014). The Delphi technique in educational research. *SAGE Open*, *4*(2), 1-8. https://doi.org/10.1177/2158244014529773

Greenbaum, T. L. (1998). *The handbook for focus group research* (2nd ed.). SAGE Publications. https://doi.org/10.4135/9781412986151.n12

Gutiérrez-Martín, A., Pinedo-González, R., & Gil-Puente, C. (2022). Competencias TIC y mediáticas del profesorado. Convergencia hacia un modelo integrado AMI-TIC. *Comunicar, 70*, 21-33. https://doi.org/10.3916/C70-2022-02

Hsu, C., & Sandford, B. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research and Evaluation*, *12*, 10. https://doi.org/10.7275/pdz9-th90

Hultgren, A. K., Owen, N., Shrestha, P., Kuteeva, M. and Mežek, Š. (2022). Assessment and English as a medium of instruction: Challenges and opportunities. *Journal of English-Medium Instruction, 1*, 105–123. https://doi.org/10.1075/jemi.21019.hul

Inbar-Lourie, O. (2022). EMI programs and formative assessment. *Journal of English-Medium Instruction*, *1*(2), 204–231. https://doi.org/10.1075/jemi.21014.inb

Kaktiņš, L. (2019). *Does Turnitin support the development of international students' academic integrity? Ethics and Education*, *14*, 430-448. https://doi.org/10.1080/17449642.2019.1660946

Kao, Y-T., & Tsou, W. (2017). EMI Course Assessment: A Survey Study of the Issues. In: Tsou, W., Kao, SM. (Eds.), *English as a Medium of Instruction in Higher Education. English Language Education*, *8*. Springer. https://doi.org/10.1007/978-981-10-4645-2_11

Khalil, M., & Er, E. (in press). Will ChatGPT get you caught? Rethinking plagiarism detection. https://doi.org/10.48550/arXiv.2302.04335

Lasagabaster, D. (2022). Teacher preparedness for English-medium instruction. *Journal of English-Medium Instruction*, *1*(1), 48-64. https://doi.org/10.1075/jemi.21011.las

Lasagabaster, D. (2018). Fostering team teaching: Mapping out a research agenda for English-medium instruction at university level. *Language Teaching*, *51*(3), 400-416. https://doi.org/10.1017/s0261444818000113

Li, J., & Li, M. (2018). Turnitin and peer review in ESL academic writing classrooms. *Language Learning & Technology*, *22*, 27–41. https://dx.doi.org/10125/44576

Li, N., & Wu, J. (2018). Exploring Assessment for Learning Practices in the EMI Classroom in the Context of Taiwanese Higher Education. *Language Education & Assessment*, *1*(1), 28–44. https://doi.org/10.29140/lea.v1n1.46

Liang, W., Yusekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). *GPT detectors are biased against non-native English writers*. arXiv. https://doi.org/10.48550/arXiv.2304.02819

Linstone, H.A., & Turoff, M. (1976). The Delphi method. Techniques and applications. *Technometrics*, *18*, 363-374. https://doi.org/10.2307/1268751

Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping out a research agenda for generative artificial intelligence in tertiary education. *Australasian Journal of Educational Technology*, *39*(1), 1-8. https://doi.org/10.14742/ajet.8695

Lund, B., & Ting, W. (in press). Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries? Library Hi Tech News. https://dx.doi.org/10.2139/ssrn.4333415

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing*, *18*(4), 351-372. https://doi.org/10.1177/026553220101800403

Macaro, E. (2022). English medium instruction: What do we know so far and what do we still need to find out? *Language Teaching*, *55*, 533-546. https://doi.org/10.1017/S0261444822000052

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Language Teaching*, *51*, 36-76. https://doi.org/10.1017/S0261444817000350

Macfarlane, B., Zhang, J., & Pun, A. (2012). Academic integrity: A review of the literature. *Studies in Higher Education*, *39*, 339-358. https://doi.org/10.1080/03075079.2012.709495

Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4354422

Milligan, L. O., & Tikly, L. (2018). *English as a medium of instruction in postcolonial contexts: Issues of quality, equity and social justice*. Routledge.

Milligan, L. O., & Tikly, L. (2016). English as a medium of instruction in postcolonial contexts: Moving the debate forward. *Comparative Education*, *52*(3), 277-280. https://doi.org/10.1080/03050068.2016.1185251

Morgan, D. L. (1997). *Focus groups as qualitative research: 16 (qualitative research methods)* (2nd ed.). SAGE Publications.

Morreel, S., Mathysen, D., & Verhoeven, V. (2023). Aye, AI! ChatGPT passes multiple-choice family medicine exam. *Medical Teacher*. https://doi.org/10.1080/0142159X.2023.2187684

Mortenson, L. (2022). Integrating social justice-oriented content into English for academic purposes (EAP) instruction: A case study. *English for Specific Purposes*, *65*, 1-14. https://doi.org/10.1016/j.esp.2021.08.002

Mortenson, L. (2021). White TESOL Instructors' Engagement with Social Justice Content in an EAP Program: Teacher Neutrality as a Tool of White Supremacy. *BC TEAL Journal*, *6*(1), 106–131. https://doi.org/10.14288/bctj.v6i1.422

Otto, A., & Estrada Chichón, J. L. (2021). Analysing EMI Assessment in Higher Education. *Revista Tempos e Espaços em Educação*, *14*, e15475. http://dx.doi.org/10.20952/revtee.v14i33.15475

Passey, D. (2019). Technology-enhanced learning: Rethinking the term, the concept, and its theoretical background. *British Journal of Educational Technology*, *50*, 972-986. https://doi.org/10.1111/bjet.12783

Readings, B. (1997). *The university in ruins*. Harvard University Press. https://doi.org/10.2307/j.ctv1cbn3kn.9

Robinson, J. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher*, *39*, 582-590. https://doi.org/10.3102/0013189X10389811

Rolfe, G. (2013). *The university in dissent*. Routledge. https://doi.org/10.4324/9780203084281

Rothbauer, P.M. (2008). Triangulation. In L.M. Given (Ed.), *The SAGE encyclopedia of qualitative research methods. Volumes 1 & 2* (pp. 892-894). SAGE Publishing. https://doi.org/10.4135/9781412963909.n468

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

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). Can AI-Generated Text be Reliably Detected? arXiv preprint arXiv:2303.11156. https://doi.org/10.48550/arXiv.2303.11156

Sadler, D.R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*, 535-550. https://doi.org/10.1080/02602930903541015

Sah, P. K. (2022). A research agenda for English-medium instruction. *Journal of English-Medium Instruction*, *1*(1), 124-136. https://doi.org/10.1075/jemi.21022.sah

Scheibe, M., Skutsch, M. & Schofer, J. (2002). Experiments in Delphi Methodology. In H. A. Linstone & M. Turoff (Eds.), *The Delphi method. Techniques and applications* (pp. 257-281). Addison Wesley Publishing Company.

Shamim, F. (2023). EMI, ELT, and Social Justice. Case of Pakistan. In R. A. Giri, A. Padwad, & M. M. N. Kabir (Eds.), *English as a Medium of Instruction in South Asia* (pp. 92-111). Routledge. https://doi.org/10.4324/9781003342373-7

Shen, Y., Heacock, L., Elias, J., Hentel, K.D., Reig, B., Shih G. & Moy, L. (in press). ChatGPT and Other Large Language Models are Double-edged Swords. *Radiology*. https://doi.org/10.1148/radiol.230163

Sterling, S., Plonsky, L., Larsson, T., Kytö, M., & Yaw, K. (2023). Introducing and illustrating the Delphi method for applied linguistics research. *Research Methods in Applied Linguistics*, *2*(1), 100040. https://doi.org/10.1016/j.rmal.2022.100040

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books. http://bit.ly/3Z2JT6Y

Tai, H-Y. (2015). Writing development in syntactic complexity, accuracy and fluency in a content and language integrated learning class. International Journal of Language and Linguistics, *2*, 149156. https://bitly.ws/UGvg

Thangaratinam, S., & Redman, C. W. E. (2005). The Delphi technique. *The Obstetrician & Gynaecologist, 7*, 120-125. https://doi.org/10.1576/toag.7.2.120.27071

UCL (February, 2023). *AI, education and assessment*. https://www.ucl.ac.uk/teaching-learning/assessment-resources/ai-education-and-assessment-staff-briefing-1

Uztosun, M. S. (2018). Professional competences to teach English at primary schools in Turkey: A Delphi study. *European Journal of Teacher Education*, *41*(4), 549-565. https://doi.org/10.1080/02619768.2018.1472569

van der Walt, C., & Kidd, M. (2013). Acknowledging academic biliteracy in higher education assessment strategies: A tale of two trials. In A. Doiz, D. Lasagabaster & J. M. Sierra (Eds.), *English-medium instruction at universities: Global challenges* (pp. 27-43). Multilingual Matters. https://doi.org/10.21832/9781847698162-006

von der Gracht, H. A. (2012). Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting and Social Change*, *79*, 1525-1536. https://doi.org/10.1016/j.techfore.2012.04.013

Peter Bannister Predoctoral Fellow at Universidad Internacional de La Rioja, member of the PRODIGI research group, BALEAP MA Dissertation Award winner (2023), who has published internationally on implications of GenAI for teaching, learning, and assessment in EMI settings through involvement in nationally funded research projects on GenAI and assessment.

Dr Alexandra Santamaría-Urbieta Associate Professor at Universidad Internacional de La Rioja, Spain and member of PRODIGI research group, who obtained PhD in Tourist Translation from University of Las Palmas de Gran Canaria, Spain. She has international publications on technological innovation in English language teaching. Principal Investigator of nationally funded research projects exploring role of GenAI in higher education assessment.

Dr Elena Alcalde-Peñalver Associate Professor at the University of Alcalá (Spain) who holds a PhD in Translation from the University of Granada (Spain) and member of the FITISPos research group. She has teaching and research experience at national and international levels in the field of foreign languages and is currently carrying out research into the implications of GenAI in the field.