AI in education

MDPI

*Article*

# Fair Marking in the Generative AI Era: Introducing the Master's Dissertation Marking Framework

**Mireilla Bikanga Ada**

School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK; mireilla.bikangaada@glasgow.ac.uk

**Abstract**

This paper presents the Master's Dissertation Marking Framework (MDMF), a longitudinally developed framework designed to support fairer and more transparent master's dissertation assessment. The framework was developed through a multi-phase, design-based research framework, comprising a literature review, a survey and in-depth interviews (2022) conducted prior to the emergence of generative AI, and follow-up empirical phases between 2023 and 2025. Across these phases, the framework evolves from an initial focus on procedural consistency and bias mitigation to a broader sociotechnical perspective that incorporates ethical boundaries, professional judgement, institutional responsibility, and the disruptive effects of generative AI on assessment practice. The paper traces the progression of the framework to MDMF Version 5, the final iteration, which consolidates six interdependent components: ethical boundaries and AI policy clarity; fairness and equity issues; pre-marking tasks and calibration; marker allocation; marking processes, culture, and well-being; and technology as both enabler and disruptor. Drawing on empirical evidence from academic staff involved in MSc dissertation marking in the post-generative-AI context, the framework brings together these components to address both longstanding and emerging challenges in assessment. The findings demonstrate that fairness in dissertation marking cannot be achieved through procedural mechanisms or technological solutions alone. Instead, the MDMF supports fairer assessment by structuring human judgement, enabling calibration, and clarifying ethical boundaries in AI-mediated contexts. The framework offers a coherent yet adaptable model for institutions seeking to maintain valid and defensible assessment practices in the age of generative AI.

**Keywords:** dissertation marking; assessment fairness; framework; postgraduate education; professional judgement; generative AI in higher education; marking and moderation



Academic Editor: Yupei Zhang

Received: 6 February 2026
Revised: 29 March 2026
Accepted: 28 May 2026
Published: 2 July 2026



## 1. Introduction

The project or master's dissertation is a central yet complex element of postgraduate study, often described as an "elusive chameleon" due to its varied formats, diverse student body, and supervision styles (Pilcher, 2011). Marking dissertations is a high-stakes assessment practice and an integral part of assessment and feedback, which is crucial to student learning and development (Bikanga Ada, 2014). The UK's Quality Assurance Agency (QAA) views assessment as fundamental to the student experience and mandates that it be fair, consistent, reliable, and valid (Quality Assurance Agency, 2018). This includes clear criteria, weighting, and moderation policies, allowing academic staff to exercise judgement fairly, specifically in group work and projects. Despite these guidelines, dissertation marking remains a complex task (Bikanga Ada, 2023). This project addresses that challenge by

developing a framework to support fair and consistent marking at the master's level. Addressing this challenge requires attention not only to procedural consistency but also to the interpretive, relational, and increasingly sociotechnical nature of dissertation assessment.

### 1.1. Background

Although assessment and feedback are often discussed together, recent scholarship cautions against conflating their purposes. Winstone and Boud (2022) argue that the entanglement of assessment and feedback can obscure the distinctive learning function of feedback while also overloading assessment processes with developmental expectations. This distinction is especially important in dissertation marking, where the present study is primarily concerned with the fairness and defensibility of summative evaluative judgement, even though such judgments may also shape students' learning and future development. However, these strands of research are often treated separately, leaving little guidance on how to integrate these perspectives into coherent dissertation-marking practices.

#### 1.1.1. Marking Reliability and Fairness

Recent evidence from a scoping review of graduate research project assessment confirms persistent concerns regarding supervisor involvement in summative marking, with supervisors often awarding higher marks than independent examiners due to familiarity, role conflict, and investment in the research process (Ernstzen et al., 2026). According to OfQual, reliability in assessment refers to the consistency of outcomes when conditions such as examiner, time, or version are varied (Tisi et al., 2013, p. 10). Factors influencing reliability include rubric design, marker behaviour, and marking processes (Tisi et al., 2013, p. 11). Double-blind marking is often recommended to reduce bias and overestimated agreement (Tisi et al., 2013, p. 7). While involving supervisors in marking can ensure subject expertise (McQuade et al., 2020), it can also introduce bias, including leniency and halo effects, often tied to interpersonal familiarity (Wolf, 2015; Vinton & Wilke, 2011; Jackson, 2018). McQuade et al. (2020) found that early-career supervisors may inflate marks due to perceived professional risk, and even internal markers unfamiliar with students were not immune to bias, although subject expertise during moderation helped mitigate this. While supervisors contribute valuable subject expertise and holistic insight into the research process, their dual role as a mentor and assessor introduces tensions between developmental support and objective evaluation, reinforcing concerns about fairness and reliability in dissertation marking (Ernstzen et al., 2026).

Fairness in marking is subjective and contextual. Gordon and Fay (2010, p. 97) suggest fairness is "in the eye of the beholder," while Burger (2017) and Gipps and Stobart (2009) critique overly technical interpretations. Some markers compare student work to check rank order and adjust accordingly (Crisp, 2013). Hudson et al. (2017) found that academics associated fairness with consistent criteria and contextual awareness, while external examiners prioritised fairness to students via local marking criteria. Pitt and Winstone (2018) found that anonymous marking did not significantly affect fairness perceptions. In contrast, McQuade et al. (2020) demonstrated that in a large postgraduate sample, internal assessors graded slightly higher than externals, with moderation reducing this gap, highlighting both systemic inconsistency and the corrective role of moderation.

#### 1.1.2. Rubrics, Exemplars and Judgement

Assessment artefacts, such as rubrics, marking schemes, and descriptors, support consistency (D. Sadler, 2014) and enable students to interpret quality (Bearman & Ajjawi, 2018). Beyond their function as evaluative tools, recent studies suggest that assessment tasks also create opportunities to develop evaluative judgement through situated interactions with peers, teachers, and disciplinary practices (Fischer et al., 2024). Exemplars, in particular,

can help anchor criteria to practice and develop evaluative judgement (Tai et al., 2018; Jönsson & Prions, 2019). When used in scaffolded and dialogic ways, rubrics and exemplars can support the development of evaluative and productive knowledge, self-monitoring, and self-regulation (Hawe et al., 2021). However, when criteria are too generic, markers experience "conceptual acrobatics" (Hudson et al., 2017) in applying them across varied disciplines. Furthermore, experienced assessors often rely on tacit, holistic judgment, using rubrics retrospectively to justify decisions rather than to generate them (D. R. Sadler, 2009; Bloxham et al., 2011). Man et al. (2020) and Chakraborty et al. (2021) found that markers use both criterion- and norm-referencing, and the effectiveness of rubrics can vary with task type, cohort, and experience. These findings suggest that while rubrics and exemplars are important, they do not fully resolve the complexity of evaluative judgment in dissertation marking.

#### 1.1.3. Training on How to Use the Rubric

Just like assessment literacy represents students' understanding of the criteria against which their performance is evaluated (Carless & Boud, 2018), marking literacy represents assessors' understanding of the criteria against which they assess student work. However, assessors often lack confidence or feel stressed (Marr & Forsyth, 2010), and confusion persists between processes such as second marking and moderation (Forsyth et al., 2015). Targeted training has been shown to improve consistency in marking, though short digital formats may have a limited impact (Davis, 2016; Pufpaff et al., 2015). Pérez-Ros et al. (2021), for example, found that trained supervisors awarded more consistent marks, while untrained assessors tended to be more generous. Similarly, Postmes et al. (2022) highlight the need for rubric-based training that focuses on key predictors of final grades and provides structured formative feedback. Importantly, effective training extends beyond procedural instruction. Evidence suggests that rubrics and exemplars are most effective when their use is explicitly scaffolded through explanation, modelling, dialogue, and guided practice, rather than relying solely on static documentation (Peeters et al., 2014; Hawe et al., 2021). This highlights the need to view marking literacy as a socially developed capability rather than an individual skill.

#### 1.1.4. Moderation and Standards

One method used for quality assurance in higher education is the moderation of assessment judgements. Moderation is defined as a process that ensures fair and consistent outcomes aligned with assessment criteria (Quality Assurance Agency, 2006). Moderation serves multiple purposes, including promoting equity, providing justification for grades, ensuring accountability, and supporting calibration among assessors (Bloxham et al., 2016). The most common are justification and accountability, particularly where grades are contested. It also offers opportunities for new or less experienced markers to develop shared understandings of standards through engagement with peers (Bloxham & Boyd, 2012). However, moderation does not guarantee accuracy and presents challenges, such as increased workload and slow feedback (Bloxham, 2009; Satchell & Pratt, 2010; Elliot et al., 2011). While moderation can reduce discrepancies between markers, it functions as a corrective mechanism rather than addressing the underlying sources of variation in evaluative judgment.

#### 1.1.5. Marking Practices and Interpretive Bias

Assessment judgments in dissertation marking are shaped by tacit knowledge, disciplinary expectations, and contextual interpretation. Bloxham and Boyd (2012) found that lecturers rarely cite pressure to inflate grades but acknowledged the vagueness of standards and the challenge of reducing complex student work to a single mark. Bourke

and Holbrook (2013) demonstrated that criteria such as English fluency, although necessary, are insufficient to determine thesis quality. Williams and Kemp (2019) warn that rigid rubrics may harm creativity and reduce the authenticity of thesis contributions. Similarly, Ashworth et al. (2010) argue that expectations must adapt to student diversity (e.g., disabilities) to maintain fairness, sometimes requiring the reinterpretation of standard rubrics. This highlights the interpretive nature of marking. Even when using the same rubric, assessors may apply criteria differently, leading to variation in marks (Ernstzen et al., 2026). Supervisors, in particular, may interpret student work through their knowledge of the research process and their relationship with the student, which can introduce both valuable insight and potential bias. Recent evidence indicates that supervisors often award higher marks than independent examiners, reflecting tensions between subject expertise, mentorship roles, and objective evaluation (Ernstzen et al., 2026).

These differences are further shaped by contextual and social factors. Assessment has been described as a socially situated practice in which judgments of quality emerge through interaction with tasks, peers, and disciplinary norms rather than through criteria alone (Fischer et al., 2024). Similarly, recent analysis of UK assessment policy suggests that the mere availability of assessment criteria does not ensure shared understanding or consistent application, particularly when socio-cultural and socio-material dimensions of transparency are underdeveloped (Gonsalves & Lin, 2025).

Overall, this body of work highlights that variability in dissertation marking is not merely a technical issue, but a consequence of the inherently interpretive and relational nature of evaluative judgment. In a mixed-methods study of master's dissertation marking in Computing Science, Bikanga Ada (2023) identified persistent challenges to fairness, noting that all assessors are susceptible to bias. Contributing factors include the growing number of students, assessors marking unfamiliar topics, inconsistent use of generic marking schemes, and the lack of anonymity in the marking process. These issues persist despite the use of exemplars, shared marking schemes, and moderation practices intended to ensure consistency. Therefore, given the challenges in achieving fairness in marking these dissertations, a Master Dissertation Marking Framework (MDMF) is necessary to help guide a fairer marking process.

#### 1.1.6. Generative AI and Feedback/Assessment

The rapid emergence of generative artificial intelligence (GenAI) has introduced new possibilities and tensions in higher education assessment and feedback. Recent studies suggest that AI systems can, in some contexts, provide feedback comparable to that of human instructors, particularly for structured tasks and when guided by appropriate prompts (Çağlar-Özhan et al., 2025). GenAI also enables more immediate, iterative, and scalable feedback processes, potentially supporting student engagement and addressing workload constraints in higher education (Zhan et al., 2025).

However, these capabilities are accompanied by important limitations. Empirical research indicates that students' perceptions of feedback are shaped not only by its content but also by the perceived identity of the provider, with AI-generated feedback often viewed as less credible or less genuine when its source is disclosed (Nazaretsky et al., 2026). In addition, emerging theoretical work emphasises that effective feedback is not purely informational but relational, grounded in processes of trust, recognition, and shared vulnerability that AI systems cannot fully replicate (Corbin et al., 2025).

These developments highlight a critical distinction between feedback and high-stakes assessment. While GenAI may support aspects of formative feedback and efficiency, its role in complex, judgment-intensive contexts such as dissertation marking remains uncertain. Dissertation assessment involves the evaluation of originality, authorship, critical

engagement, and disciplinary standards—processes that rely on contextual interpretation and professional judgment (Bearman et al., 2024; Williamson & Eynon, 2020). Moreover, emerging evidence suggests that AI-supported educational processes may reproduce or reshape bias rather than remove it, including uneven trust in AI-generated outputs and biases linked to perceived credibility. More broadly, bias in generative AI can arise from training data, model design, and deployment contexts, reinforcing the need for transparency, monitoring, and human oversight in high-stakes applications (Afreen et al., 2025). These challenges highlight the deeply contextual and relational nature of dissertation assessment, raising fundamental questions about whether such judgment-intensive processes can be meaningfully supported, rather than oversimplified or distorted, by AI systems.

### *1.2. Study Rationale and Research Questions*

Despite extensive research on assessment, feedback, and rubric use, limited work has integrated these insights into a coherent, empirically grounded framework for master's dissertation marking, particularly in the context of generative AI. Existing research has largely focused on either traditional assessment challenges or emerging AI-enabled feedback practices in isolation, with insufficient attention to how these intersect in high-stakes, judgment-based assessment contexts.

This paper reports on a multi-phase, Design-Based Research-informed study that develops and refines a framework for dissertation marking through iterative empirical investigation. The study traces the development of the Master's Dissertation Marking Framework (MDMF) across multiple phases and examines how it responds to both long-standing assessment challenges and emerging AI-related disruptions. Consistent with this approach, this study adopts a sociotechnical perspective on assessment, conceptualising dissertation marking as an interaction between human judgment, institutional practices, and technological mediation. It is further informed by Design-Based Research (DBR) and grounded in scholarship on assessment validity, evaluative judgment, and feedback, positioning marking as a complex and relational practice. This study is guided by the following research questions:

- RQ1: What challenges affect fairness, consistency, and judgment in master's dissertation marking?
- RQ2: How do these challenges evolve in the context of generative AI?
- RQ3: How can a structured, empirically grounded framework support fairer and more defensible dissertation assessment practices?

These research questions guide the iterative development and refinement of the MDMF across all phases of the study.

## 2. Research Design and Methodology

This section outlines the research design and methodology adopted to develop and iteratively refine the Master's Dissertation Marking Framework (MDMF).

### *2.1. Research Design: A Design-Based Research Approach*

This study adopts a Design-Based Research (DBR) approach (Armstrong et al., 2018) to develop and refine the Master's Dissertation Marking Framework (MDMF) across multiple iterative phases conducted between 2022 and 2025, spanning pre- and post-generative-AI adoption contexts. DBR is particularly suited to addressing complex, practice-based challenges in authentic educational contexts, where solutions must be both theoretically informed and empirically grounded. It enables the iterative design, testing, and refinement of interventions through close engagement with practitioners and real-world settings.

In the context of dissertation assessment, fairness, consistency, and validity are not purely technical problems but are shaped by disciplinary norms, institutional practices, and professional judgment. The emergence of generative AI further complicates these dynamics by introducing new uncertainties around authorship, originality, and evaluative criteria. A DBR approach, therefore, allows the framework to evolve responsively, incorporating both longstanding assessment challenges and emerging AI-related disruptions.

The framework development follows a cyclical approach based on the Generic Model for Design Research (GMDR) proposed by McKenney and Reeves (2012). Figure 1 presents the GMDR model as adapted by (Bikanga Ada, 2018), comprising iterative cycles of (a) Analysis and exploration (identifying problems and needs), (b) Design and construction (developing the framework), and (c) Evaluation and reflection (refining the framework based on empirical evidence).

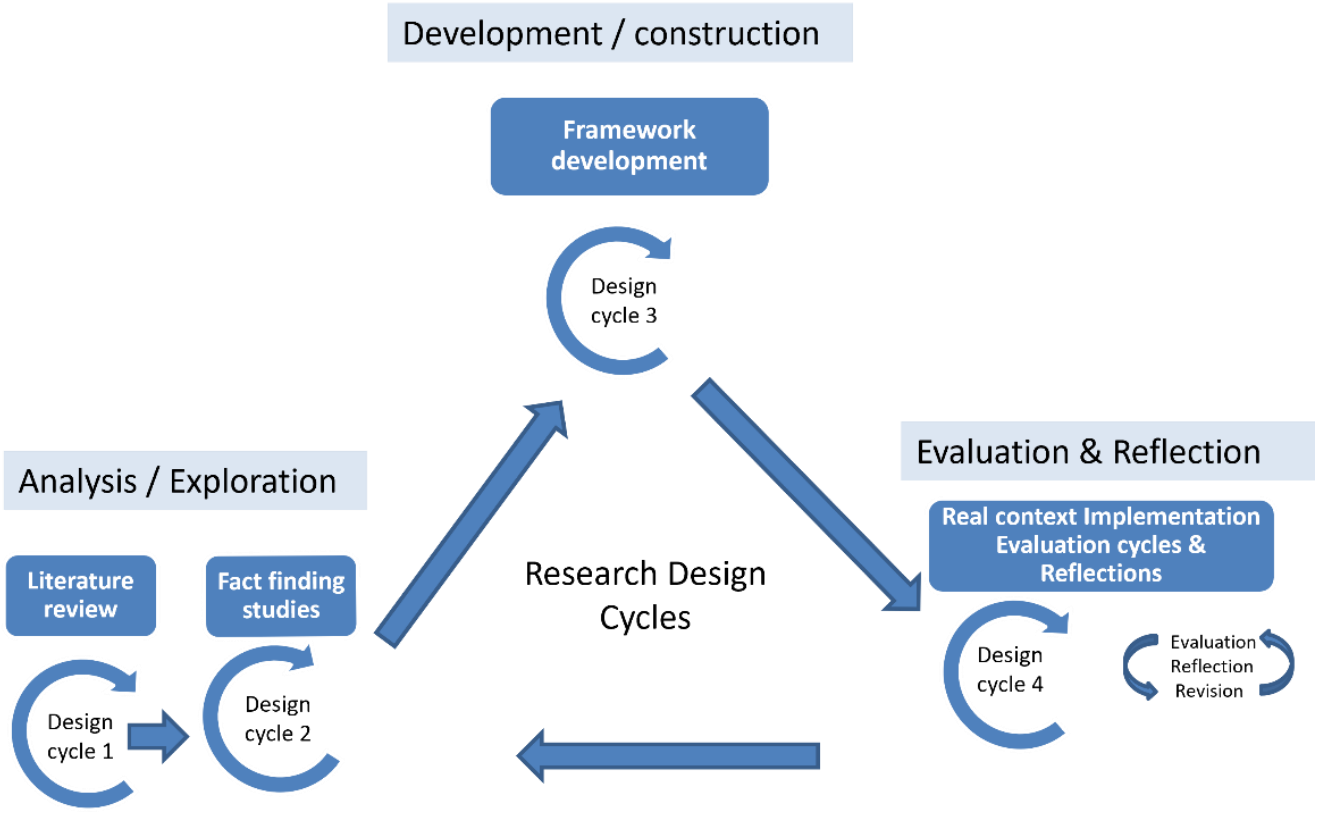


**Figure 1.** Design-based research cycles adopted from McKenney and Reeves (2012) and Bikanga Ada (2018).

Overview of Iterative Research Phases

The development of the MDMF was conducted through six interrelated phases, each contributing to successive iterations of the framework. Table 1 summarises the design, participants, methods, and outcomes of each phase. Collectively, these phases address the research questions as follows: Phases 1–3 identify challenges in dissertation marking (RQ1), Phases 4–6 examine how these challenges evolve in the context of generative AI (RQ2), and all phases contribute to the iterative development and refinement of the MDMF (RQ3).

**Table 1.** Overview of DBR Phases in the Development of MDMF.

| Phase | Year | Purpose | Method | Participants | Analysis | Framework Output |
|---|---|---|---|---|---|---|
| Phase 1 | 2022 | Identify key issues in dissertation marking | Literature review | — | Narrative synthesis | MDMF v1 |
| Phase 2 | 2022 | Explore staff experiences and challenges | Survey | $n$ = 31 academic staff | Descriptive + thematic analysis | MDMF v2 |
| Phase 3 | 2022 | Deepen understanding of marking practices | Semi-structured interviews | $n$ = 7 academic staff | Thematic analysis | MDMF v3 |
| Phase 4 | 2023 | Capture early responses to generative AI | Two exploratory surveys | $n$ = 19; $n$ = 8 academic staff | Descriptive analysis | Exploratory insights informing AI-related framework adaptation |
| Phase 5 | Early 2025 | Examine post-AI marking practices | Semi-structured interviews | $n$ = 3 academic staff | Reflexive thematic analysis | MDMF v4 |
| Phase 6 | Late 2025 | Evaluate AI-related challenges and perceptions | Mixed-methods survey | $n$ = 13 (quantitative); $n$ = 14 (qualitative) | Descriptive statistics + qualitative analysis | MDMF v5 |

### *2.2. Participants*

All empirical phases were conducted within a UK higher education institution, focusing on academic staff involved in Master's dissertation supervision and marking, primarily in Computing Science and related disciplines. Participants across phases included dissertation supervisors, second markers (readers), and academic staff involved in moderation. They had varying levels of experience in dissertation assessments, ranging from early-career academics to experienced staff with extensive marking responsibilities. This diversity enabled the study to capture a broad range of perspectives on fairness, marking practices, and the evolving impact of generative AI.

### *2.3. Data Collection Methods*

Literature review (Phase 1): An initial literature review was conducted to identify key challenges in dissertation marking, including fairness, reliability, rubric use, moderation, and assessor judgment. This informed the development of the first version of the framework (MDMF v1).

Surveys (Phases 2, 4, 6): Multiple surveys were used to capture staff perceptions across different stages:

- Phase 2 (2022). A survey of 31 academic staff explored challenges in dissertation marking, including fairness, bias, and marking practices.
- Phase 4 (2023). Two exploratory surveys ($n$ = 19; $n$ = 8) captured early responses to the emergence of generative AI prior to the establishment of institutional policies.
- Phase 6 (2025). A mixed-methods survey collected both quantitative data (Likert-scale items on AI effectiveness, concerns, and support) and qualitative data (open-ended responses on marking challenges and AI impact).

Interviews (Phases 3 and 5): Semi-structured interviews were conducted to gain deeper insights into marking practices and evolving challenges:

- Phase 3 (2022): Seven academic staff participated in interviews exploring marking culture, fairness, bias, and the use of marking schemes.

- Phase 5 (2025): Three follow-up interviews were conducted with experienced markers after at least one full dissertation marking cycle in the post-generative-AI context. These interviews focused on changes in marking practices, perceptions of AI, and emerging fairness concerns.

### 2.4. Data Analysis

Different analytical approaches were used across phases, aligned with the exploratory and iterative nature of DBR: Qualitative data (interviews and open-ended survey responses) were analysed using thematic approaches. In Phase 5, reflexive thematic analysis (Braun & Clarke, 2006) was employed to identify patterns in participants' experiences and perceptions. Quantitative data (Phase 6) were analysed descriptively (means, standard deviations, medians), given the small, purposive sample. Composite subscales were constructed to summarise perceptions of AI effectiveness, concerns, and support. Non-parametric tests were used cautiously to explore patterns rather than support generalisation. The purpose of analysis across all phases was not to produce statistically generalisable findings, but to generate design-relevant insights that informed iterative refinement of the framework.

### 2.5. Framework Development and Iterative Refinement

This iterative refinement process addresses RQ3 by translating empirical findings into successive versions of the framework and design decisions. At each phase, findings were used to refine the MDMF:

- MDMF v1: Derived from literature (conceptual foundation)
- MDMF v2–v3: Refined using staff perceptions (survey and interviews)
- MDMF v4: Updated to incorporate AI-related challenges emerging from early post-ChatGPT 3.5 and 4 contexts
- MDMF v5: Final refinement based on mixed-methods evidence from the post-generative-AI environment

Rather than introducing entirely new components at each stage, the framework evolved through extension of existing components, re-weighting of priorities, and incorporation of emerging challenges (e.g., AI-mediated fairness issues). All phases were conducted with institutional ethical approval (approval numbers anonymised for review). Participation was voluntary, and responses were anonymised. Given the small sample sizes and the identifiable institutional context, care was taken to ensure the confidentiality of qualitative data reporting.

## 3. Master's Dissertation Marking Framework Design Cycles in 2022

This section reconstructs the development of the framework prior to the public release of generative AI (including ChatGPT) and explains the context surrounding each iteration of the MDMF.

### 3.1. MDMF Version 1—Initial Draft Based on the Literature Review

The first draft of the Master's Dissertation Marking Framework (MDMF) was developed from the literature review and corresponds to the analysis and exploration phases of the adapted Generic Model for Design Research (GMDR). The review revealed a lack of a framework for marking master's dissertations, despite extensive discussion of individual components such as rubrics, moderation, and assessor roles. While these elements are widely recognised as supporting fairness, persistent challenges remain, particularly in the consistent application of criteria and the role of assessor judgment. This initial iteration therefore synthesises these insights into a structured framework, providing a conceptual starting point for subsequent empirical refinement. The literature highlights multiple components of dissertation assessment

but does not integrate them into a coherent framework. Despite the widespread use of marking criteria and rubrics, persistent challenges related to fairness and consistency remain. This first cycle, therefore, consolidates these insights into an initial structured solution, forming the foundation of the MDMF. MDMF considers three main components (Figure 2).

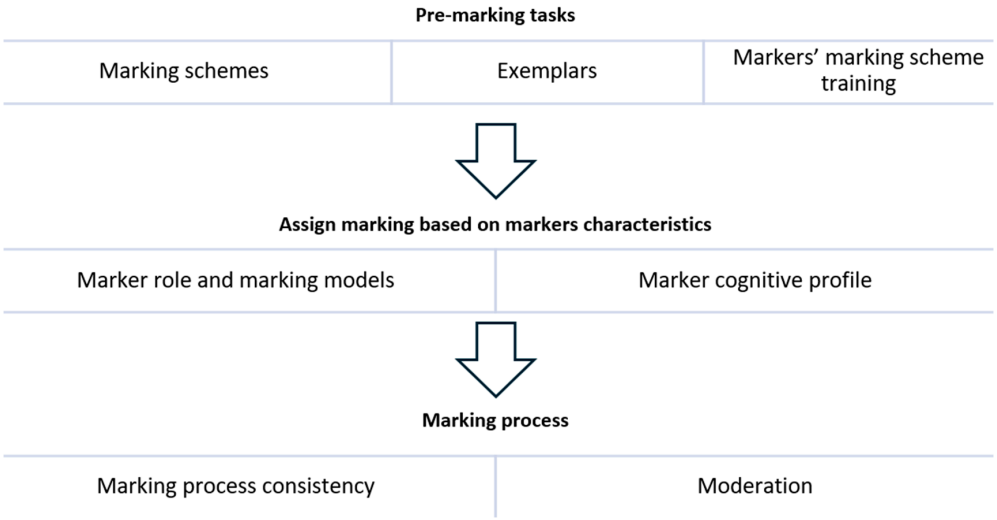


**Figure 2.** Master's Dissertation Marking Framework (MDMF) based on literature.

**(1) The pre-marking tasks**, which involve creating appropriate marking schemes, providing exemplars and assessors taking training on using the marking scheme, which is an important tool for marking dissertations, including ensuring fairness (Chakraborty et al., 2021; Hudson et al., 2017; Peeters et al., 2014; Pérez-Ros et al., 2021; D. Sadler, 2014; Tai et al., 2018; Williams & Kemp, 2019).

**(2) Assigning marking based on markers' characteristics**. In this phase, an effort should be made to consider the role of markers, the marking models, and the cognitive profiles of markers (McQuade et al., 2020). Indeed, during marking, the supervisors' knowledge of their students' backgrounds put the second marker in an unfair position, as they had to mark based on what was presented to them. It is even more noticeable when they assess dissertations on unfamiliar topics. However, the second marker's most obvious unfair practice is their strategic approach to marking when they know who the supervisors are. They adapt their marking, thus no longer reflecting the objectivity required to ensure fairness in the marking process. This could be why many participants would prefer complete anonymity in the marking process to ensure they are not 'observed' (Bikanga Ada, 2023). However, anonymity does not always ensure fairness (Pitt & Winstone, 2018).

**(3) Marking process**, which requires finding ways to ensure consistency and moderation, as the different moderation forms can be confusing (Forsyth et al., 2015).

### *3.2. MDMF Version 2—Framework Updates Based on Fact-Finding Study 1 (Survey Conducted in 2022)*

The second draft of the Master's Dissertation Marking Framework (MDMF), Figure 3, draws on fact-finding study 1, a survey of 31 participant teaching staff (Bikanga Ada, 2023), and aligns with the GMDR model. Findings from fact-finding study 1 (survey) (Ethical Approval Number: 402210101, 13 December 2021) agree with the literature and present detailed issues related to the master's dissertation marking process in the School of CS (see context). The original core components, (1) Pre-marking tasks, (2) Assign marking and (3) Marking process, have not changed and now contain detailed information on the solutions as suggested by participants. Two additional core components have been added to the framework: (4) Identifying issues related to fairness and equity in master's dissertation

marking. This is the first step of the framework to be completed. It focuses on understanding the marking culture and guides the proposals and implementation of the solutions. These issues are related to the project dissertation, marker characteristics, the marking process, marking schemes, the technology used for marking and the impact of these issues on the marker's well-being (Bikanga Ada, 2023). The second new core component is (5) Technology, which plays a critical role as the facilitator of all changes and implementations. With the use of technology, assessors can improve and refine their marking practices. For example, it can help capture projects and assessors' characteristics, ensuring that correct mappings are implemented. Furthermore, technology could help ensure the anonymity of marking, thus making assessors feel less "observed" (Bikanga Ada, 2023).

### *3.3. MDMF Version 3—Framework Updates Based on Fact-Finding Study 2 (Interviews Conducted in 2022)*

This section proposes a revised draft of the Master's Dissertation Marking Framework (MDMF) (Figure 4), which draws on fact-finding study 2 (interviews conducted in 2022, prior to the introduction of ChatGPT). Findings from fact-finding study 2 (interviews, $n = 7$) agree with the literature and study 1 (survey) (Ethical Approval Number: 402210101, 13 December 2021).

The core components remain the same. However, marking culture, which includes discussions and debates around dissertation marking, has been added to the core component (3) Marking process. Creating a marking culture is important because it could help avoid assessors being "anxious" and "worked up" during the marking process, resulting in a negative marking experience. One example is stress and anxiety caused by intercultural differences when marking dissertations at a new institution for the first time. Markers want blind double and third marking as well as blind negotiation during moderation. However, it is necessary to recognise that where a completely blind marking process is in place, this might remove the need for open debate and discussions during the marking process. In core component (4) "Identifying issues related to fairness and Equity in master's dissertation marking", comparing dissertations is added as an issue under "marking process" because, under the current marking policy, dissertations should not be compared. Moreover, although markers sometimes use comparisons between students' work to fine-tune marks and check their rank order to ensure fairness (Crisp, 2013), the complexity of the topics and the levels of markers' expertise and internal bias could render any comparison unfair. Finally, language bias, identified by two of the three non-native assessors interviewed, has also been included as an issue under "marker characteristics". Language bias is another element that affects fairness. Educational institutions are multicultural and have been embracing internationalisation. This has led to an increase in recruiting staff and students for whom English is not their primary language. The study (Bikanga Ada, 2023) shows that non-native markers can be more lenient on a poorly written dissertation. This could be because they unconsciously associate their own limitations in the native language with those of students who wrote the dissertation they are marking. Based on their own experience and the literature, one non-native interviewee notes the existence of "cognitive dissonance" during the marking process and argues that dissertations should not be penalised based on poor language use toward culturally diverse students with limited English. This aligns with cognitive dissonance theory (Festinger, 1957), where inconsistencies between expectations and observed performance may influence the evaluative judgment (Harper, 2016). Indeed, it seems that when marking, some native speakers may have unconscious thoughts that non-native speakers are less intelligent because their dissertations are not well written. In a way, it could be said that native speakers unconsciously disassociate themselves from the characteristics of non-native students as non-native speakers and adhere to the objectivity of their own beliefs.

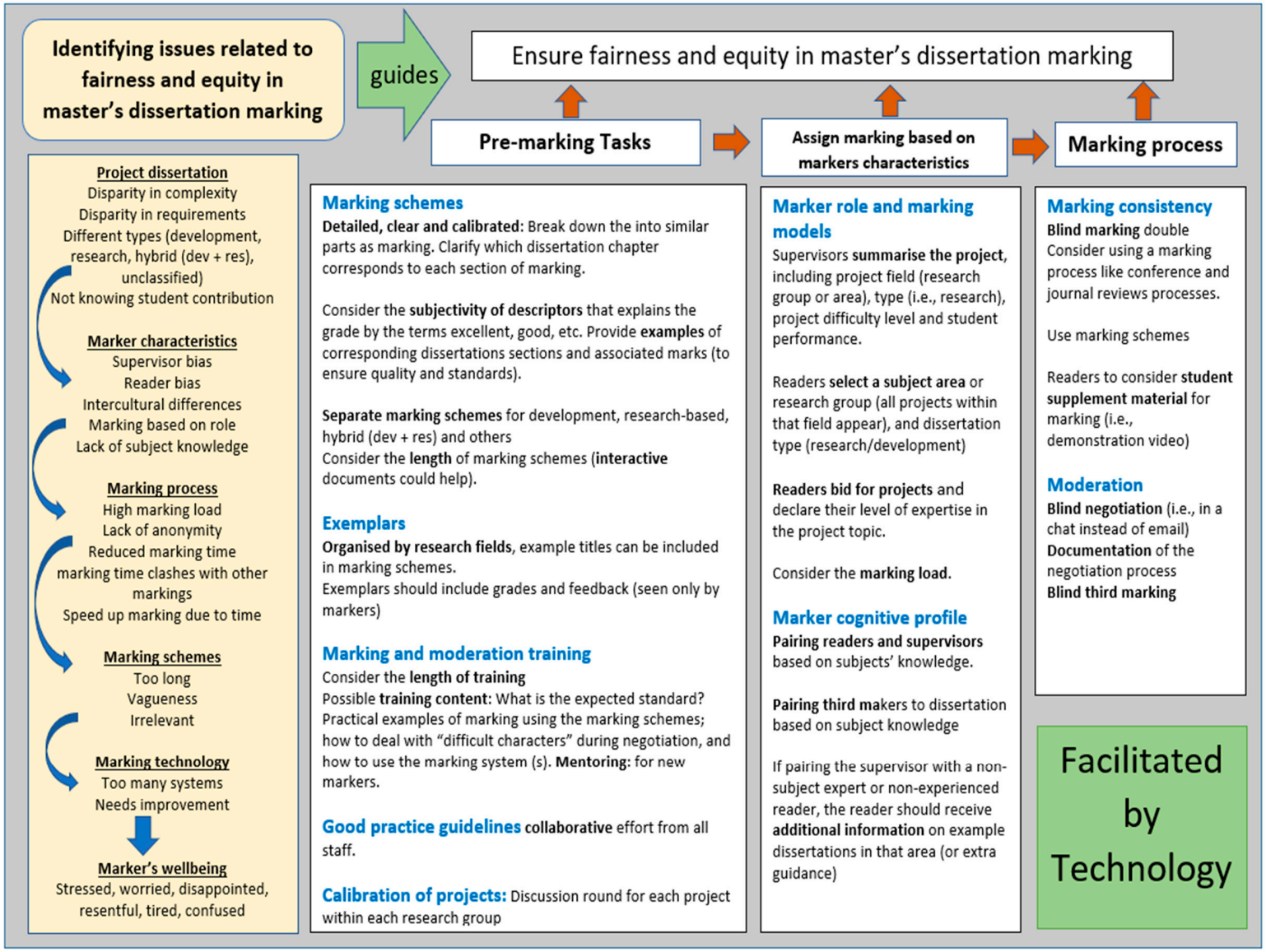


**Figure 3.** Master's Dissertation Marking Framework (MDMF) version 2, based on literature and survey 2022.

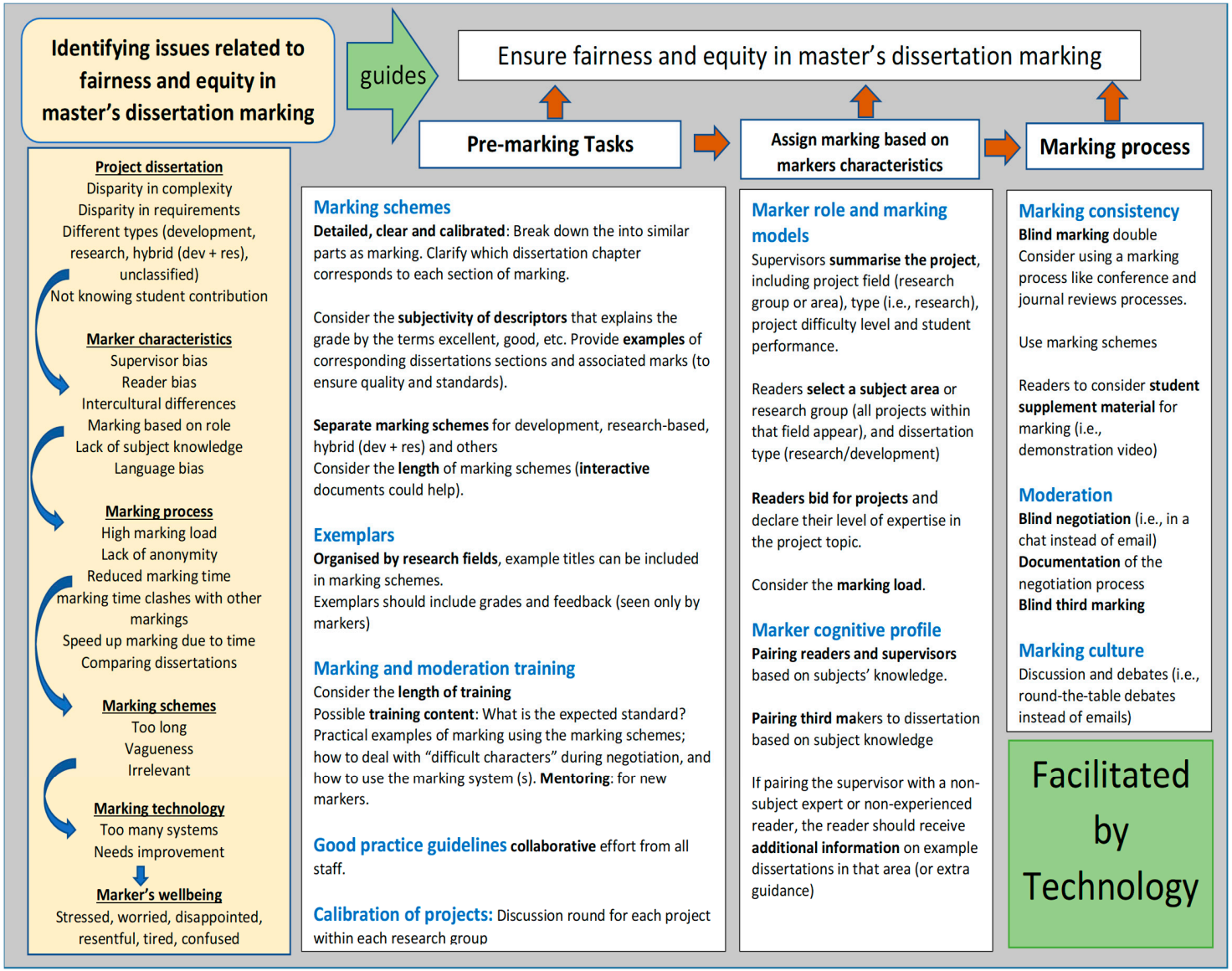


**Figure 4.** Updated Master's Dissertation Marking Framework (MDMF) version 3 based on literature, survey and interviews (2022).

## 4. Transitional Phase Evidence: Early Staff Responses to Generative AI in Dissertation Marking in 2023

Before the framework could be further evaluated in practice, the public release of ChatGPT in late 2022 introduced an immediate disruption to academic assessment. To capture early staff responses during this transitional phase—prior to the establishment of institutional policies, shared norms, or stabilised marking practices—two exploratory surveys were conducted in 2023 (Ethical Approval Numbers 300220170, 13 June 2023; 300230028, 18 October 2023). The first was administered in mid-2023 (from June), during the initial period of generative AI exposure ($n$ = 19), and the second in October 2023, after staff had completed MSc dissertation marking for that academic year ($n$ = 8). Across the two surveys, responses were obtained from academic staff with varied disciplinary backgrounds and marking experience. While awareness of ChatGPT was high, direct use of generative AI for dissertation marking was extremely limited. In the mid-2023 survey, 16 of 19 respondents (84%) reported never using ChatGPT for any aspect of dissertation marking, and only 4 indicated awareness of its possible use for marking. Similarly, in the post-marking survey conducted in October 2023, most respondents (5 out of 8) reported no use of ChatGPT in marking, and 3 did not answer the question.

By contrast, personal or pedagogical use of ChatGPT was common, particularly for drafting text, generating examples, refining language, or exploring ideas. However, respondents consistently distinguished these uses from evaluative judgment. Across both surveys, dissertation marking was framed as a professional responsibility requiring disciplinary expertise, contextual interpretation, and accountability. Attitudes toward AI-assisted assessment were predominantly sceptical. In response to the question of whether ChatGPT could make dissertation marking more fair or objective, most respondents selected "No" or "Strongly Disagree," with only a small minority selecting "Maybe," and none expressing clear agreement. Descriptive ratings of AI effectiveness were consistently low across core assessment challenges, including supervisor bias, reader bias, lack of anonymity, and uncertainty about student contribution. Conversely, ratings of concern were high, particularly regarding AI's limited ability to evaluate originality, difficulties aligning AI outputs with human marking criteria, risks of bias, lack of contextual understanding, and the potential for over-dependence on automated systems. Although a small number of respondents identified narrowly defined, peripheral roles for AI, such as summarising text or supporting understanding of unfamiliar technical content, these were explicitly framed as supportive rather than decisional. Support for AI replacing existing dissertation marking practices was uniformly low across both surveys.

This transitional-phase evidence indicates that ethical resistance, boundary-setting, and the defence of professional judgment emerged immediately after the introduction of generative AI, rather than developing gradually over time. Although this early evidence did not prompt structural changes to the framework prior to Version 4, it provided a critical empirical signal: generative AI was already being understood by markers less as a solution to fairness or workload challenges and more as a source of new ethical, epistemic, and procedural risk. Consequently, these findings served as a diagnostic checkpoint rather than a basis for redesign, indicating that further empirical evidence grounded in post-dissertation marking practice was required before substantive framework modification. These findings served as a diagnostic checkpoint, indicating that assessors positioned generative AI not as a solution but as a source of new ethical, epistemic, and procedural risks. This required further investigation grounded in actual marking practice, which informed the subsequent phases of framework refinement.

While this transitional-phase evidence clarified early ethical positioning and resistance to AI-assisted assessment, it did not capture how dissertation marking practices evolved

once generative AI became embedded within completed marking cycles, necessitating a subsequent, practice-based investigation.

## 5. Post-Marking Practice Evidence in 2025: Understanding Shifts in Dissertation Marking in the AI-Dominated Era

A subsequent phase of empirical investigation was required to examine how dissertation marking practices evolved once AI became embedded in assessment contexts. To address this, a follow-up qualitative study was conducted using in-depth interviews with three former participants, focusing on their experiences marking master's dissertations during the 2023–2024 academic year (Ethical Approval number: 300240064, 10 January 2025). The interviews took place in January and February 2025, after participants had completed at least one full cycle of MSc dissertation marking in the post-ChatGPT context. Reflexive thematic analysis was conducted following Braun and Clarke's (2006) six-phase approach, ensuring that the analysis remained grounded in participants' accounts and that illustrative quotations were used accurately and representatively.

### *5.1. Results from the Early 2025 Interview Data*

#### 5.1.1. Continuities in Marking Practices and Persistent Challenges

Despite rapid technological and institutional changes, core challenges identified in 2022 remain prominent in 2025. Subjectivity in dissertation marking continues to be influenced by supervisor familiarity, emotional connection, perceptions of student effort, and prior involvement, all of which impact the objectivity of the marker. As Staff Z explained, "... as a supervisor, you'll know the student because you have been interacting with them...," underlining the unavoidable pastoral dimension of dissertation supervision that may influence marking. Despite some improvements, including a clearer distinction between research and development projects, the marking scheme remains vague; markers still employ informal calibration strategies (e.g., reading all allocated dissertations before final grading) to mitigate their subjective judgment and improve consistency and perceived fairness.

#### 5.1.2. Post-ChatGPT Shifts: AI Awareness, Ethical Ambiguities, and Reframing Fairness

The emergence and widespread adoption of generative AI tools have introduced new considerations into the dissertation marking process. Markers now actively assess whether text may have been generated by AI, citing signs such as a uniform tone, absence of voice, or fabricated references as red flags. As Staff U noted, "I sometimes have doubts in terms of ... like pieces of text that look very much not just translated but generated." (Staff U).

Participants repeatedly called for explicit, student-facing policies on acceptable AI use. They had a shared concern about the ambiguity between acceptable academic support (e.g., grammar checking or language enhancement) and potential misconduct (e.g., using AI to generate entire paragraphs or arguments). This ambiguity, if unaddressed, risks confusing both students and assessors and undermining shared expectations around academic integrity. This aligns with Perkins and Roe (2024), who found that many academic integrity policies lacked clear language on AI, highlighting an urgent need for "Technological Explicitness" to distinguish AI from misconduct.

Staff U suggested practical solutions, such as checklists, to help students navigate expectations. However, Staff V argued, "... there is no existing generative AI that is ethically acceptable." This stricter stance challenges assumptions that regulation alone can address deeper ethical concerns. Indeed, high-level principles and regulations may appear promising but lack proven methods to translate principles into practice... and robust legal and professional accountability mechanisms (Mittelstadt, 2019).

#### 5.1.3. AI's Disruptive Impact on the Originality and Academic Integrity of the Dissertation

The use of ChatGPT and other generative AI tools is reshaping traditional understandings of originality, authorship, and creativity in academic work (Mei et al., 2025), particularly in dissertation projects. Participants stressed that critical thinking and independent analysis, rather than mere novelty or surface-level originality, should define academic originality. Staff V raised a concern about the practical implications of using AI to assess originality, noting that they would not be able to tell the difference between someone who has used ChatGPT and someone who has not, "assuming they both know how to write good English." Similarly, as Staff U affirmed, critical engagement is the core value of the dissertation: "The point of the dissertation is to have the students work on their . . . critical thinking skills."

#### 5.1.4. Resistance to Using AI Tools for Marking Dissertations

Beyond policy ambiguity, all three participants explicitly rejected the use of generative AI for direct dissertation marking, citing ethical risks, loss of professional integrity, and reinforcement of linguistic or stylistic bias. Staff Z described the use of AI for direct evaluation or feedback as "lazy" and "unethical," arguing it undermines the trust and effort invested by both students and supervisors. Marking, they emphasised, is a human process rooted in reciprocity and professional integrity, values diminished when replaced by automation. He stated he would be "pretty cross" if his own work were assessed this way. Staff Z also cautioned against unfairly harsh judgments based on suspicion about AI, urging markers to self-reflect before penalising students. Additionally, he noted the emotional toll of repetitive AI-generated content, describing a growing sense of "fatigue" from reviewing similar-sounding texts.

Staff V viewed the use of ChatGPT in marking as unethical and a threat to the core values of higher education. In contrast, Staff U took a more reflective stance, acknowledging that in-house AI could assist as a third marker or provide structured feedback. However, they cautioned that AI lacks contextual insight, such as knowledge of a student's academic journey, and raised concerns about confidentiality, cultural bias, and fairness, particularly the risk of disadvantaging students whose writing does not reflect dominant linguistic norms.

#### 5.1.5. Fairness Reframed: From Procedural Consistency to AI-Mediated Equity

Staff V pursues fairness through benchmarking to ensure internal consistency but acknowledges the complexity of this approach. They note that subject expertise can unintentionally lead to harsher judgments, which they consider unfair compared with marking unfamiliar topics. In discussing AI use, Staff Z likened it to students copying from Stack Overflow, an academic risk, but not grounds for punishment without evidence. They emphasised a relational model of fairness, suggesting that supervisors, due to their close engagement with students, are better positioned to assess authorship and originality fairly. These accounts suggest that fairness is no longer understood only as procedural consistency between markers, but increasingly as an equity issue shaped by AI use, authorship uncertainty, and differential access to linguistic and technological support.

#### 5.1.6. Language Bias Reinterpreted in an AI Context

Language bias, already noted in 2022, has taken on new dimensions with the advent of AI-enhanced fluency. Participants expressed concern that AI-generated language could mask weak content or give some students an unfair advantage, as fluent English, especially when assisted by AI, can distort perceptions of quality. Staff V avoids penalising non-native features that do not hinder meaning but admits that emotional reactions to poor English

can influence marking. They also highlighted equity concerns, where students with better English skills or access to AI tools may appear more competent, potentially disadvantaging others. This reflects an evolving form of language bias influenced by the use of AI.

### *5.2. MDMF Version 4—Framework Updates Based on Early 2025 Study*

MDMF Version 4 (Figure 5) builds on earlier framework iterations by incorporating the main themes arising from the early 2025 follow-up interviews, particularly ethical boundary-setting, AI-related ambiguity in authorship and originality, evolving language bias, and the growing emotional burden of post-ChatGPT marking. The six core components are further elaborated in Tables 2–7.

#### 5.2.1. Core Component 1: Ethical Boundaries and Policy Clarity in the Age of AI

This newly added core component, detailed in Table 2, addresses the pressing need for shared institutional definitions and expectations regarding the use of AI tools in project dissertations. It functions as the ethical and procedural backbone of the entire framework. Without clear boundaries and policies, fairness cannot be upheld, integrity risks go unresolved (Gonsalves, 2024), and markers' judgment becomes inconsistent or overly cautious. Participants' comments highlight that fairness now depends on clarity, transparency, and consistency of AI policy enforcement. Participants identified several pressing concerns, including the lack of institutional policy or guidance on acceptable AI use (Petricini et al., 2025), confusion about what constitutes misconduct versus legitimate support, students' non-disclosure of AI-use declarations (Gonsalves, 2024), markers' uncertainty on how to act on suspected but undeclared AI-generated content, difficulty evaluating fluent but AI-generated work in terms of originality or critical thinking, anxiety around ethical limits of student and staff use of AI; no consistent norms across subject areas on AI-allowed practices; unclear distribution of responsibility between institutions, markers, and students; risk of unintentionally penalising students unfamiliar with AI tools or unaware of expectations; concerns that AI access may amplify educational privilege and language-based inequities; lack of formal procedures when AI use is suspected or contested.

These findings indicate an urgent institutional responsibility to co-create clear, inclusive, and enforceable policies (An et al., 2025) supported by transparent processes for declaration, assessment, and dispute resolution. Without this component, academic integrity and equitable marking in the age of AI cannot be sustained.

#### 5.2.2. Core Component 2: Identifying Issues Related to Fairness and Equity in Master's Dissertation Marking

Core component 2 (see Table 3) maintains the original structure while incorporating AI-specific concerns, reframed biases, and newly emerging tensions. It also includes a new sub-component, Institutional Policy & Ethics.

#### 5.2.3. Core Component 3: Pre-Marking Tasks

Table 4 presents the updated pre-marking preparation tasks. Participants emphasised that fairness now depends not only on procedural consistency but also on clarity, transparency, and institutional alignment around AI-related expectations (Chai et al., 2024). As a result, training must now include guidance on how to identify and assess AI-assisted writing. Jakesch et al. (2023) found that people often fail to recognise AI-generated text, relying on flawed cues, such as first-person pronouns, contractions, or familiar topics, to judge whether writing is human. These cues can be deliberately mimicked, making AI-generated content appear even more human than human-written content.

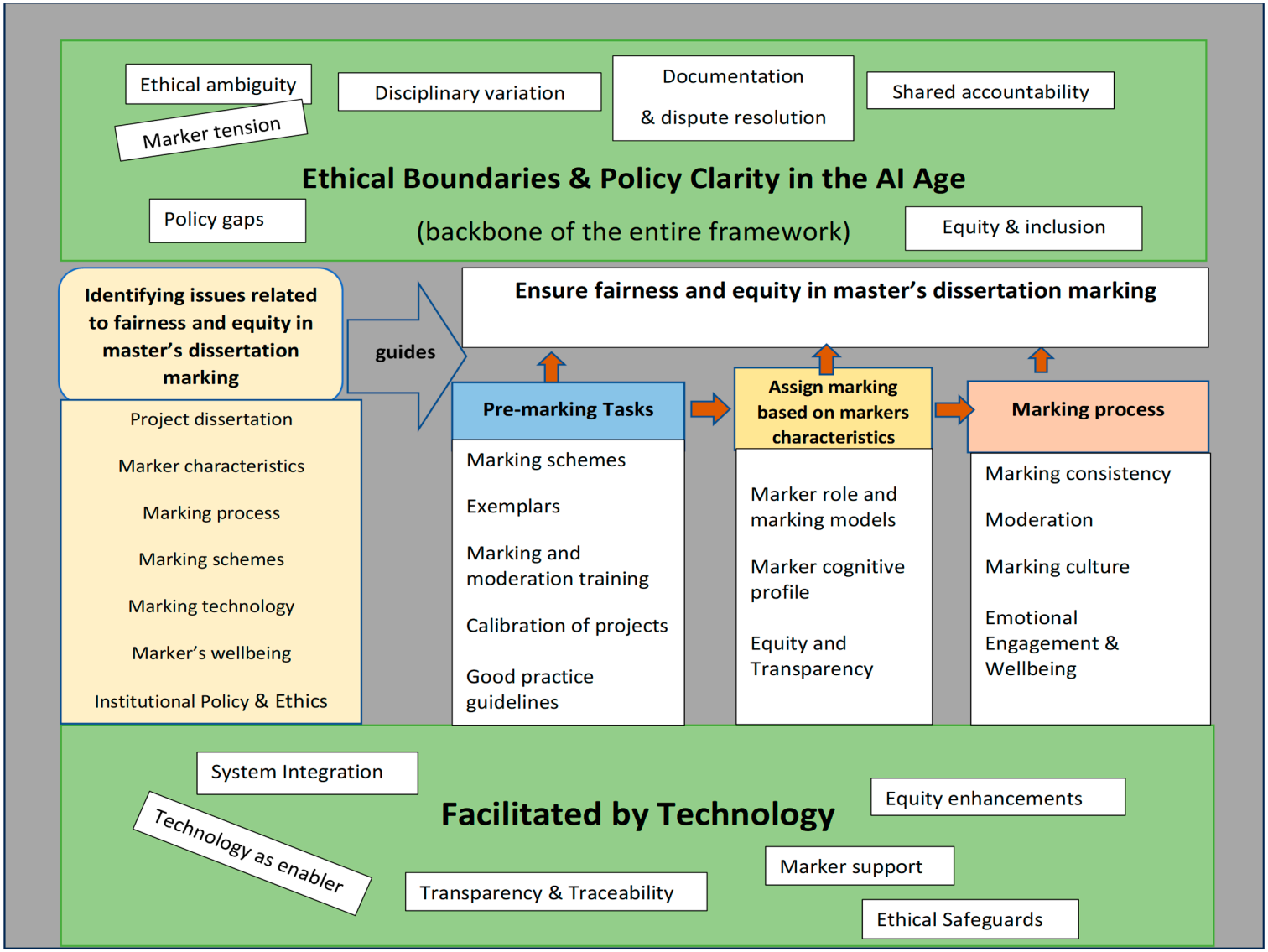


**Figure 5.** Updated Master's Dissertation Marking Framework (MDMF) version 4, based on the post-AI study data (early 2025).

**Table 2.** Ethical boundaries and policy clarity in the age of AI (New).

| Sub-Components | Elements |
| --- | --- |
| Policy gaps | Co-develop clear, student-facing AI use policies<br>Distinguish acceptable vs. unacceptable uses<br>(e.g., grammar vs. argument generation)<br>Align policies with existing integrity frameworks. |
| Ethical ambiguity | Introduce AI use declarations as a standard for dissertations<br>Provide marker guidance on how to handle undeclared or suspicious<br>AI use scenarios. |
| Marker tension | Develop AI-sensitivity checklists for markers.<br>Foster reflective discussions on the boundaries of responsible AI use in<br>research writing. |
| Disciplinary variation | Encourage departments to develop discipline-specific AI guidelines that align with<br>broader policy.<br>Use these to inform marking schemes and moderation approaches. |
| Shared accountability | Promote a shared responsibility model: Institutions provide guidance, markers<br>exercise judgment, and students disclose their AI use transparently. |
| Equity and Inclusion | Include equity lens in policy-making (e.g., ensure access to AI tools is considered)<br>Support multilingual and neurodiverse students in understanding AI guidelines. |
| Documentation and Dispute Resolution | Develop transparent documentation protocols for suspected AI use.<br>Include processes for student appeals, staff consultation, and third-party<br>review if needed. |

**Table 3.** Identifying issues related to fairness and equity in master's dissertation marking (* denotes post-ChatGPT additions).

| Sub-Components | Elements |
| --- | --- |
| Project Dissertation | Disparity in complexity<br>Disparity in requirements<br>Different types (development, research, hybrid (dev + res), unclassified)<br>Not knowing the student contribution<br>AI-assisted work further obscures student contribution *<br>Difficulty in judging independent thinking and authenticity without<br>explicit declaration * |
| Marker Characteristics | Supervisor bias<br>Reader bias<br>Intercultural differences<br>Marking based on role<br>Lack of subject knowledge<br>Language bias<br>Fluency bias amplified by AI tools *<br>Greater cognitive dissonance due to AI-mediated language quality *<br>Emotional reaction to AI-sounding text * |
| Marking Process | High marking load<br>Lack of anonymity<br>Reduced marking time<br>Marking time clashes with other markings<br>Speed up marking due to time constraints<br>Comparing dissertations<br>Fatigue from reviewing similar-sounding AI-generated work *<br>Suspicion of AI use without evidence *<br>Emotional disengagement from similar language patterns * |

**Table 3.** *Cont.*

| Sub-Components | Elements |
| --- | --- |
| Marking Schemes | Too long<br>Vagueness<br>Irrelevant<br>Not adapted to assess AI-mediated work *<br>Lack of guidance for handling AI involvement *<br>No differentiation between AI-generated and human voice * |
| Marking Technology | Too many systems<br>Needs Improvement<br>No integrated AI-detection tools *<br>Risk of data misuse (e.g., student work used to train AI) *<br>No tools to support transparency or traceability of AI use * |
| Marker Well-being | Stressed, worried, disappointed, resentful, tired, confused<br>Increased emotional fatigue from AI-authored text *<br>Moral discomfort marking suspected AI work *<br>Frustration at policy voids and shifting norms * |
| Institutional Policy & Ethics (cross-cutting issue linked to Component 1) * | (New in V4—For full detail, see Component 6—Ethical Boundaries and Policy Clarity)<br>No consensus on acceptable AI use *<br>Tension between innovation and integrity *<br>Need for clearer policy and student-facing guidelines * |

**Table 4.** Pre-marking Tasks (* denotes post-ChatGPT additions).

| Sub-Components | Elements |
| --- | --- |
| Marking Schemes | Detailed, clear and calibrated: Break down the project into similar parts as marking. Clarify which dissertation chapter corresponds to each section of marking.<br>Consider the subjectivity of descriptors that explain the grade using terms such as 'excellent,' 'good,' etc. Provide examples of corresponding dissertation sections and associated marks (to ensure quality and standards).<br>Separate marking schemes for development, research-based, hybrid (dev + res) and others<br>Consider the length of marking schemes (interactive documents could help).<br>Explicit AI-related criteria or guidance (e.g., evaluating AI-assisted work) *<br>Examples of how AI use affects structure or tone *<br>Clarify how originality and critical thinking should be assessed in the AI era * |
| Exemplars | Organised by research fields, example titles can be included in marking schemes.<br>Exemplars should include grades and feedback (seen only by markers)<br>Exemplars could include AI-assisted and non-AI-assisted versions (for comparison)*<br>Tag exemplars with critical thinking quality, not just polish or fluency * |
| Training for Marking and Moderation | Consider the length of training.<br>Possible training content: What is the expected standard?<br>Practical examples of marking using the marking schemes; how to deal with "difficult characters" during negotiation, and how to use the marking system (s).<br>Mentoring: for new markers.<br>Training to detect possible signs of AI-generated content (e.g., voice consistency, fabricated references) *<br>Discussions of bias and fairness in AI-mediated language *<br>Handling undeclared AI use in line with ethical policy * |
| Calibration Practices | Discussion round for each project within each research group<br>Include AI-awareness calibration, e.g., how AI usage affects language, criticality, and structure *<br>Share edge cases (e.g., suspicious fluency vs. weak reasoning) * |
| Good Practice Guidelines | Collaborative authorship by all marking staff<br>Updated regularly to reflect institutional AI policy *<br>Include decision-making flowcharts for suspected or declared AI involvement * |

#### 5.2.4. Core Component 4: Assigning Marking Based on Marker Characteristics

The updates for this component are detailed in Table 5 and include a new sub-component, Equity and Transparency.

**Table 5.** Core Component 4—Assigning Marking Based on Marker Characteristics (* denotes post-ChatGPT additions).

| Sub-Components | Elements |
| --- | --- |
| Marker Roles & Marking Models | Supervisors summarise the project, including project field (research group or area), type (i.e., research), project difficulty level and student performance.<br>Readers select a subject area or research group (all projects within that field appear), and dissertation type (research/development).<br>Readers bid for projects and declare their level of expertise in the project topic. Consider the marking load.<br>Add AI expertise as part of expertise declaration (e.g., ability to detect or handle AI-generated work) *<br>Factor in emotional load when assigning AI-heavy disciplines or tasks *—Emotional implications of sustained AI exposure are further addressed under Emotional Engagement in Table 5: Marking Process. |
| Marker cognitive profile | Pairing readers and supervisors based on subjects' knowledge.<br>Pairing third markers to dissertations based on subject knowledge.<br>If pairing the supervisor with a non-subject expert or non-experienced reader, the reader should receive additional information on example dissertations in that area (or extra guidance).<br>If not paired with a subject expert, provide extended support to include examples of AI-influenced dissertations (to help identify patterns). *<br>Address language-bias risk by balancing native and non-native speakers across pairings (where necessary). * |
| Equity and Transparency (New) * | Prevent hidden bias against AI-generated fluency (e.g., polished but superficial content). *<br>Include reflective calibration sessions to align perceptions of AI use and authenticity. *<br>See also Marking Culture in Table 5: MDMF V4—Marking Process for practical procedures to support alignment during assessment. |

#### 5.2.5. Core Component 5: Marking Process

As seen in Table 6, a strong "marking culture" remains central, but it must now adapt to the integration of AI. For example, there are concerns that markers may unconsciously penalise or suspect AI use even when it is absent, creating equity challenges. The table also includes a new sub-component: Emotional Engagement & Well-being.

**Table 6.** Core Component 5: Marking Process (* denotes post-ChatGPT additions).

| Sub-Components | Elements |
| --- | --- |
| Marking Consistency | Blind marking, double marking.<br>Consider using a marking process, such as conference and journal review processes<br>Use marking schemes.<br>Readers to consider student supplement material for marking (i.e., demonstration video)<br>Consider AI detection uncertainty: introduce flagging protocols for suspected AI use. *<br>Ensure schemes clearly define originality, criticality, and engagement in AI-mediated texts. * |

**Table 6.** *Cont.*

| Sub-Components | Elements |
| --- | --- |
| Moderation | Blind negotiation (i.e., in a chat instead of email)<br>Documentation of the negotiation process<br>Blind third marking: If necessary, the third marker should receive AI-context training on how to assess suspected generated content fairly. *<br>Moderation discussions include AI use declarations or absences. * |
| Marking Culture | Discussion and debates (i.e., round-the-table debates instead of emails)<br>Create space for AI-focused calibration discussions (e.g., signs of AI use, thresholds for concern). For example, use anonymised samples and structured discussions to align marker judgments on AI use, especially regarding fluency vs. criticality. *—see Table 4: Equity and Transparency for assessor alignment practices during marker assignment.<br>Address marker fatigue or frustration with repeated AI-generated language. *<br>Reflective calibration sessions to address and align perceptions of AI use and authorship. *<br>Prevent hidden bias against polished but superficial AI-generated fluency. * |
| Emotional Engagement & Well-being (new) * | Recognise fatigue from marking similar-sounding or impersonal AI-generated text. *<br>Offer institutional debriefs or informal spaces for assessors to reflect on emotionally or ethically challenging marking situations. *<br>(See also Table 4: Marker Roles and Cognitive profile for how emotional well-being is considered during the allocation of AI-heavy projects.) |

#### 5.2.6. Core Component 6: Technology

As detailed in Table 7, this core component now encompasses not only technology's facilitative role but also its disruptive influence. Institutional (in-house) systems must protect student data and avoid embedding linguistic or cultural biases into marking algorithms.

**Table 7.** Core Component 6—Technology (* denotes post-ChatGPT additions).

| Sub-Components | Elements |
| --- | --- |
| Technology as Enabler | Use platforms to capture project/assessor characteristics<br>Aid mapping of markers to dissertations<br>Improve anonymity<br>Integrate AI-awareness features, e.g., space for AI use declarations *<br>Add dashboards to flag patterns in submission styles or language that may suggest AI involvement * |
| Systems Integration | Too many systems<br>Push for unified platforms that allow end-to-end marking, moderation, and communication. *<br>Consider carefully governed tools that support human review of possible AI involvement, while recognising the limitations of automated detection. * |
| Transparency & Traceability | Store and log AI-use declarations, any moderation discussions around AI suspicion, and actions taken. *<br>Allow audit trails for disputed cases (e.g., if misconduct is suspected but not confirmed) * |
| Ethical Safeguards | Avoid training AI models on student data without consent *<br>Clarify data retention, privacy, and consent processes *<br>Avoid building in biases that privilege fluency or stylistic uniformity * |
| Equity Enhancements | Use tech to monitor whether markers are unintentionally favouring AI-assisted fluency. *<br>Highlight discrepancies between content quality and linguistic style as a potential indicator of equity. * |
| Marker Support | Use platforms to track emotional load or marking volume per assessor. *<br>Enable markers to flag dissertations that require support (e.g., suspected AI overuse, unclear authorship, confusing topic complexity) * |

## 6. Academic Staff Perceptions Regarding the Potential Role of AI Tools in the MSc Dissertation Marking Process at the End of 2025

This final study was conducted at the end of 2025 to capture academic staff perspectives after several years of marking master's dissertations in an AI-dominated academic environment (Ethical Approval number: 300240064, 10 January 2025). A mixed-methods survey was distributed to marking staff at the same institution, focusing on experiences of dissertation marking during the 2024–2025 academic period. Qualitative data were collected through two open-ended questions: "What challenges did you face in marking MSc dissertations this year, and how much influence did AI tools (e.g., ChatGPT or similar) have on your marking process?" and "In your view, how could the current dissertation marking process be improved?" Quantitative items asked participants to rate, on a scale from 1 (Not effective at all) to 7 (Extremely effective), the perceived effectiveness of AI tools in addressing issues identified in the dissertation marking process. Additional items asked participants to indicate their level of agreement (1 = Strongly disagree to 7 = Strongly agree) with statements relating to concerns and anticipated limitations associated with AI use in marking. Descriptive statistics for these items are reported in Table 8. Selected items were aggregated into four composite subscales: (1) effectiveness of AI tools in addressing bias-related issues, (2) effectiveness of AI tools in addressing workload and process-related issues, (3) concerns or limitations regarding AI tools, and (4) perceived support for using AI tools in dissertation marking. Descriptive statistics for these subscales are presented in Table 9.

**Table 8.** Descriptive statistics of individual survey items.

| | Item | *n* | Median | Mean | SD |
|---|---|---|---|---|---|
| | Project complexity | 13 | 1.00 | 1.92 | 1.605 |
| Bias1 | Supervisor's bias | 13 | 1.00 | 2.00 | 1.581 |
| Bias2 | Reader's bias | 12 | 1.00 | 1.83 | 1.528 |
| Bias3 | Markers' intercultural differences | 13 | 1.00 | 2.31 | 1.932 |
| Bias4 | Language bias | 13 | 1.00 | 2.00 | 1.581 |
| Wkld1 | Comparing dissertations | 13 | 1.00 | 2.46 | 1.984 |
| Wkld2 | Disparity in project requirements | 13 | 1.00 | 1.85 | 1.281 |
| Wkld3 | Lack of subject knowledge | 13 | 4.00 | 3.23 | 2.006 |
| Wkld4 | High marking load | 13 | 3.00 | 3.31 | 2.463 |
| Wkld5 | Time spent on marking | 13 | 3.00 | 3.38 | 2.293 |
| Wkld6 | Emotional and mental issues (stress, resentment, tiredness, confusion, disappointment) | 13 | 1.00 | 1.85 | 1.463 |
| Wkld7 | Vagueness or irrelevance of marking schemes | 13 | 1.00 | 1.62 | 1.193 |
| Wkld8 | Long marking schemes | 13 | 1.00 | 1.38 | 0.870 |
| Wkld9 | Diversity of project topics/areas | 13 | 1.00 | 2.15 | 1.519 |
| Wkld10 | Lack of anonymity | 13 | 1.00 | 1.46 | 1.391 |
| Wkld11 | Not knowing the student contribution | 13 | 1.00 | 1.31 | 0.751 |
| Wkld12 | Dissertation marking time clashes with other marking | 13 | 1.00 | 2.38 | 1.710 |
| conc1 | I am concerned about AI tools' limited ability to evaluate originality in a student's dissertation. | 11 | 7.00 | 6.45 | 0.688 |
| Conc1 | AI models might have difficulties aligning their grading with established human rubrics. | 11 | 7.00 | 5.09 | 2.386 |
| Conc3 | I am concerned about potential biases in AI when grading dissertations. | 11 | 7.00 | 6.09 | 1.514 |
| Conc4 | There is a risk of becoming overly dependent on AI for tasks like dissertation marking. | 11 | 7.00 | 6.64 | 0.674 |

**Table 8.** *Cont.*

| | Item | *n* | Median | Mean | SD |
|---|---|---|---|---|---|
| Conc5 | I believe the lack of personal touch in feedback from AI tools like ChatGPT is a limitation. | 11 | 7.00 | 5.09 | 2.508 |
| Conc6 | I am concerned about potential ethical issues, such as student data privacy, when using ChatGPT in the dissertation marking process. | 11 | 7.00 | 5.64 | 1.859 |
| Conc7 | The lack of context understanding in AI is a limitation for using ChatGPT in the dissertation marking process. | 11 | 7.00 | 5.36 | 2.378 |
| Conc8 | ChatGPT's lack of understanding of context hampers its effectiveness in marking dissertations. | 11 | 5.00 | 4.45 | 2.734 |
| | I have encountered other concerns/limitations, or I foresee other concerns/limitations. (Individual item) | 10 | 3.50 | 4.00 | 2.749 |
| | ChatGPT and similar technologies should replace current practices in the dissertation marking process going forward. (Individual item) | 11 | 1.00 | 1.27 | 0.647 |
| Sup1 | I could be satisfied with the quality of support and dissertation feedback provided by ChatGPT. | 10 | 1.50 | 2.50 | 1.900 |
| Sup2 | ChatGPT and similar technologies should complement human markers in the dissertation marking process. | 11 | 3.00 | 2.73 | 1.849 |
| Sup3 | Using ChatGPT can make the dissertation marking process more fair and objective. | 11 | 1.00 | 2.09 | 1.700 |

**Table 9.** Descriptives of subscales.

| | *n* | Mean | SD | Cronbach's Alpha |
|---|---|---|---|---|
| Effectiveness of AI tools in addressing bias issues identified in the dissertation marking process | 13 | 2.08 | 1.47 | 0.885 (items = 4) |
| Effectiveness of AI tools in addressing the workload/process issues identified in the dissertation marking process | 13 | 2.20 | 0.94 | 0.810 (items = 12) |
| Concerns or limitations encountered or foreseen when using ChatGPT in the dissertation marking process | 11 | 5.60 | 1.41 | 0.858 (items = 8) |
| Support encountered or foreseen when using ChatGPT in the dissertation marking process | 10 | 2.48 | 1.41 | 0.671 (items = 3) |

### *6.1. Results*

Fourteen academic staff responded to the open-ended survey questions (Lect1–Lect14). Quantitative Likert-scale items were completed by $n = 13$ for most AI "effectiveness" items, with reduced valid Ns for Concerns and Support items due to missing data ($n = 10$–$11$). Given the exploratory, framework-development focus of this study and the small, purposive sample of experienced dissertation markers, quantitative findings are interpreted descriptively and analytically rather than inferentially. Nonparametric tests are reported cautiously to identify relative patterns across constructs, not to support population-level generalisation. Rather than presenting qualitative and quantitative findings as parallel strands, results are organised around substantive dimensions of the dissertation, marking fairness and AI-related disruption, explicitly aligned with the evolving MDMF Version 5

components. Quantitative findings are used to prioritise, contextualise, and triangulate qualitative themes in support of framework refinement.

#### 6.1.1. Persistent Structural Challenges in Dissertation Marking

(MDMF v5 Components 2 and 5: Fairness & Equity; Marking Process.)

Across responses, participants consistently described structural and workload-related challenges in MSc dissertation marking that predated generative AI and have persisted in its aftermath. The most frequently cited challenges included high marking load, time pressure, diversity and complexity of project topics, and difficulty assessing individual student contribution, particularly in large cohorts. Lect2 described "the enormous marking load" as the dominant problem, while Lect7 similarly identified "the quantity of dissertations to mark" as the primary challenge. Lect11 highlighted the difficulty of acting as a reader across a "diversity/broadness of topics," and Lect12 characterised the overall process as "broken."

Descriptive statistics corroborate these accounts. Items related to workload and process received among the highest mean ratings for problem severity, including time spent on marking ($M$ = 3.38, SD = 2.29), high marking load ($M$ = 3.31, SD = 2.46), and lack of subject knowledge when acting as a reader ($M$ = 3.23, SD = 2.01). In contrast, respondents rated AI tools as largely ineffective in addressing these foundational challenges, including project complexity ($M$ = 1.92, SD = 1.61), not knowing student contribution ($M$ = 1.31, SD = 0.75), and lack of anonymity ($M$ = 1.46, SD = 1.39).

Implication for MDMF v5: These findings reinforce that fairness risks remain primarily structural and procedural, and that AI tools are not perceived as viable solutions to core workload or complexity pressures. MDMF v5 therefore retains a strong emphasis on process design, workload management, and human calibration rather than technological substitution.

#### 6.1.2. Fairness, Bias, and AI-Mediated Distortions of Judgment

(MDMF v5 Component 2: Identifying Fairness and Equity Issues.)

Concerns about fairness and bias emerged strongly across both qualitative and quantitative data. Participants repeatedly described how AI-assisted fluency complicates judgments of quality, originality, and student effort, often masking weak underlying work and compressing grades toward a perceived "average" standard. Lect5 articulated this tension clearly:

> "*I feel like projects all tend towards a B by old standards due to ubiquitous LLM use. . . all are filtered through generic competence. I want to give students who visibly did not use LLMs more credit, even if their work is less well argued, because I know it was their thoughts.*"

Similarly, Lect8 described how AI-enhanced language could "dress up an otherwise poor project," creating disagreement during moderation when readers lacked insight into the project process: "This meant having to argue the reader down on occasion, when they did not have insight into the project process."

Across responses, markers implicitly described calibration as a necessary but increasingly fragile mechanism for maintaining fairness in MSc dissertation marking in the post-generative-AI context. Several participants reported misalignment between supervisors and readers when AI-polished language obscured weaknesses in the underlying project process, requiring informal and sometimes contentious negotiation to reconcile judgments. Lect8 described having to "argue the reader down on occasion" when the reader lacked insight into the project process and was influenced by an AI-enhanced presentation. Other participants pointed to structural solutions—such as clearer rubrics, shared marking platforms, shorter and more standardised artefacts, and improved allocation of readers—

as ways to reduce interpretive divergence and support more consistent assessment. For example, Lect3 recommended shorter videos and the use of a Moodle rubric rather than the existing system, while Lect11 emphasised better templates, shorter reports, and more sensible reader allocations. Notably, although some markers used AI tools for contextual understanding of unfamiliar topics, they refrained from using AI for evaluative judgment or detection due to ethical concerns and policy constraints. Together, these findings indicate that calibration currently occurs through human negotiation and process design rather than technological arbitration, and that AI-mediated fluency has intensified the need for explicit, shared interpretive alignment among markers.

Quantitative ratings align with these concerns. Respondents perceived AI tools as having low effectiveness in addressing bias-related issues, including supervisor bias ($M$ = 2.00, SD = 1.58), reader bias ($M$ = 1.83, SD = 1.53), language bias ($M$ = 2.00, SD = 1.58), and markers' intercultural differences ($M$ = 2.31, SD = 1.93). The total bias-effectiveness subscale score was low ($M$ = 8.00, SD = 5.55).

Spearman correlation analyses further illustrate how fairness perceptions intersect with broader attitudes toward AI. Perceived effectiveness of AI tools in addressing bias was strongly positively correlated with perceived effectiveness in addressing workload/process issues ($\rho = 0.683$, $p = 0.010$) and with overall support for AI tools ($\rho = 0.914$, $p < 0.001$). In contrast, overall AI-related concerns were negatively associated with AI support ($\rho = -0.742$, $p = 0.009$).

Implication for MDMF v5: Fairness threats in dissertation marking now explicitly include AI-mediated grade distortion, fluency masking weak work, and compression of performance differentiation, extending earlier concerns about obscured contribution and authenticity.

#### 6.1.3. Ethical Resistance and Boundary-Setting Around AI Use

(MDMF v5 Component 1: Ethical Boundaries and AI Policy Clarity.)

A striking and consistent finding was the ethical resistance to using AI tools for evaluative marking. Many participants framed dissertation assessment as an inherently human, professional responsibility, rejecting the legitimacy of AI-assisted grading. Several participants explicitly stated they did not use AI tools in marking (e.g., Lect1; Lect2; Lect3; Lect4; Lect7; Lect10; Lect14). Lect6 was unequivocal: "I would never use AI tools to assess student work because I do not trust their academic judgement over my own." Lect5 similarly argued that markers should "take intellectual responsibility for this activity," questioning why AI should interpret vague marking schemes better than experienced academics.

Quantitative findings strongly corroborate these views. Respondents reported very high concern regarding AI's limited ability to evaluate originality ($M$ = 6.45, SD = 0.69), risk of over-dependence on AI ($M$ = 6.64, SD = 0.67), potential bias in AI grading ($M$ = 6.09, SD = 1.51), and ethical issues such as data privacy ($M$ = 5.64, SD = 1.86). In contrast, support for replacing current marking practices with AI was extremely low ($M$ = 1.27, SD = 0.65). A Friedman test comparing the four subscales (bias effectiveness, workload effectiveness, AI concerns, AI support) indicated a statistically significant difference in rank ordering, $\chi^2(3) = 11.73$, $p = 0.008$ ($n$ = 11). Post-hoc Wilcoxon tests showed that AI concerns were rated significantly higher than both perceived effectiveness for bias issues ($Z = -2.49$, $p = 0.013$) and workload/process issues ($Z = -2.85$, $p = 0.004$).

Implication for MDMF v5: Component 1 is reinforced as the boundary-setting backbone of the framework. Staff views indicate that institutional AI policy must clearly define what AI must not do (grading, judgment), while clarifying tightly constrained, non-evaluative uses.

#### 6.1.4. Conditional and Peripheral Roles for AI

(MDMF v5 Components 3 and 6: Pre-marking Tasks; Technology.)

Although resistance to AI-based marking was strong, some participants articulated conditional, peripheral roles for AI tools that stop short of evaluative judgment. These included summarising text, identifying consistency, and supporting understanding of unfamiliar technical content or code. Lect11 reported using AI tools to "get some context" for unfamiliar topics, while explicitly rejecting their use for assessment. Lect2 similarly noted: "They can't be used to generate actual grades. They are (potentially) useful for summarising, identifying consistency, etc." Quantitative findings reflect this ambivalence. Agreement that AI tools could complement human markers was modest ($M$ = 2.73, SD = 1.85), and satisfaction with AI-generated feedback was low ($M$ = 2.50, SD = 1.90), though still higher than support for full replacement.

Implication for MDMF v5: AI is positioned not as an assessor, but as a tightly bound infrastructural aid, reinforcing the framework's emphasis on human judgment supported, rather than replaced, by technology.

#### 6.1.5. Infrastructure, Systems, and Marking Efficiency

(MDMF v5 Component 6: Technology and Systems Integration.)

Participants also highlighted how administrative and technological infrastructure directly shapes marking efficiency and fairness. Lect1 described difficulty "trying to locate projects," recommending fewer submission points and mandatory supervisor-completed metadata. Lect3 suggested "shorter videos" and improved rubric integration within Moodle. These concerns align with quantitative evidence showing that AI tools were not perceived as effective solutions to infrastructure-related issues such as anonymity ($M$ = 1.46, SD = 1.39) or clarity of student contribution ($M$ = 1.31, SD = 0.75).

Implication for MDMF v5: Component 6 is extended to foreground submission infrastructure quality (e.g., single-point access, integrated rubrics, mandatory contextual metadata) as a core fairness-enabling condition, rather than treating technology solely as an assessment enhancement tool.

These integrated findings demonstrate that AI has intensified existing fairness tensions in dissertation marking rather than resolved them. Markers consistently resist AI-based evaluation while acknowledging limited, peripheral uses. These results directly inform MDMF Version 5 by reinforcing ethical boundary-setting, recalibrating fairness risks, and operationalising human-centred marking processes in AI-mediated academic environments.

### *6.2. MDMF Version 5—Framework Updates Based on Mixed-Methods Evidence from the Post-Generative-AI Context (Late 2025)*

MDMF Version 5 builds directly on Version 4 by incorporating empirical evidence from a mixed-methods staff survey conducted in 2025, capturing academic staff experiences of MSc dissertation marking in the post-generative-AI era. While Version 4 introduced AI as a structural disruptor to fairness, marking processes, and institutional policy, Version 5, the final version, consolidates and refines these insights by foregrounding how markers actually respond to AI in practice, ethically, cognitively, and procedurally.

Crucially, Version 5 does not add new core components, but re-weights and operationalises existing ones, reflecting strong convergence between qualitative accounts and quantitative patterns. The framework now places greater emphasis on ethical boundary-setting, AI-mediated grade distortion, human-centred calibration, and infrastructural design, while reaffirming the central role of professional judgment in dissertation assessment. MDMF v5 does not introduce new structural components but represents a

phase of theoretical consolidation, in which earlier anticipatory design elements are empirically stabilised, re-weighted, and operationalised through post-AI marking evidence. Tables 10–15 present only refinements, re-weightings, and conceptual clarifications introduced in MDMF version 5. Elements unchanged from Version 4 are intentionally omitted to emphasise theoretical progression.

#### 6.2.1. Core Component 1 (Strengthened): Ethical Boundaries and Policy Clarity in the Age of AI

In MDMF Version 5, Ethical Boundaries and Policy Clarity move from a necessary contextual component to the normative anchor of the entire framework. Empirical findings show that markers overwhelmingly resist using AI tools for evaluative judgment and view dissertation marking as an irreducible academic responsibility rather than a task to be automated.

Qualitative data reveal strong ethical positioning, with staff explicitly rejecting AI-based grading on grounds of intellectual responsibility, lack of academic judgment, and moral discomfort. Quantitatively, this resistance is reflected in very high concern scores regarding AI's ability to evaluate originality, the risk of over-dependence on AI, and potential biases in AI-mediated grading. These concerns were statistically more prominent than perceived AI effectiveness for either bias reduction or workload relief.

MDMF v5, therefore, reframes this component as the boundary-setting mechanism that determines what AI must not do in dissertation marking, rather than focusing primarily on permissible uses. The framework emphasises institutional responsibility for defining non-delegable academic judgments, clarifying staff roles, and protecting professional autonomy. Without explicit ethical boundaries, fairness, consistency, and trust in assessment decisions cannot be sustained.

Table 10 summarises the refinements to Core Component 1 in MDMF Version 5. Version 5 repositions ethical boundaries and policy clarity as the normative anchor of the framework, reflecting empirical evidence that dissertation grading is viewed by staff as a non-delegable academic judgment. The focus shifts from regulating AI use to clearly defining what AI must not do, reinforcing marker autonomy, professional responsibility, and institutional protection of human judgment in AI-mediated assessment.

#### 6.2.2. Core Component 2 (Refined): Identifying Fairness and Equity Issues in Master's Dissertation Marking

MDMF Version 5 refines the fairness and equity component by explicitly recognising AI-mediated grade distortion as a central threat. While Version 4 identified obscured student contributions and uncertainty about authenticity, Version 5 extends this to include grade compression, fluency masking weak work, and inflated perceptions of quality driven by AI-generated language. Staff accounts consistently describe difficulty distinguishing genuine intellectual contribution from AI-polished text, with several noting that dissertations increasingly cluster around similar grades, reducing meaningful differentiation. This concern is empirically reinforced by low ratings of AI effectiveness in addressing traditional bias issues (e.g., supervisor bias, reader bias, language bias), indicating that AI is not perceived as a fairness solution and may, in fact, introduce new inequities.

Version 5, therefore, reframes fairness risks as emergent properties of AI-mediated writing practices, rather than as problems solvable through technological intervention. Equity concerns now explicitly include the risk that the privileges of polished AI-assisted writing can overshadow critical engagement, disadvantaging students who rely less on AI or whose work reflects authentic but less fluent intellectual development.

**Table 10.** Core component 1 (Revised)—Ethical Boundaries and Policy Clarity in the Age of AI—Key Refinements in MDMF v5.

| Focus Area | MDMF v5 Update |
|---|---|
| Role of AI | Explicitly defines dissertation grading as a non-delegable academic judgment |
| Ethical stance | Shifts from regulating AI use to defining what AI must not do |
| Marker autonomy | Reinforces markers' professional responsibility and intellectual ownership of assessment decisions |
| Policy emphasis | Moves from permissive guidance to boundary-setting and protection of human judgment |
| Empirical grounding | Informed by strong staff resistance to AI-based grading and high ethical concern scores |

Table 11 outlines the refinements to Core Component 2 in MDMF Version 5. Drawing on mixed-methods evidence, Version 5 reframes fairness and equity risks as emergent properties of AI-mediated writing practices, including grade compression, fluency masking, and obscured student contribution. Rather than positioning AI as a mechanism for reducing bias, the framework shifts the emphasis toward human calibration, contextual judgment, and systemic equity considerations.

**Table 11.** Core component 2 (Revised)—Identifying issues related to fairness and equity in master's dissertation marking—New Emphases in MDMF v5.

| Area | MDMF v5 Refinement—New Issue: Fairness and Equity Risks |
|---|---|
| AI-mediated writing | Introduces grade compression and fluency masking as fairness risks |
| Student contribution | Reframes obscured authorship as a systemic equity issue, not an individual misconduct problem |
| Language bias | Extends from linguistic disadvantage to AI-amplified fluency bias |
| Fairness logic | Moves away from AI as bias-mitigation toward human calibration and contextual judgment |

#### 6.2.3. Core Component 3 (Operationalised): Pre-Marking Tasks and Calibration in the AI Era

While MDMF v4 expanded pre-marking tasks to include AI awareness, Version 5 further operationalises this component by foregrounding human-led calibration as the primary mechanism for maintaining standards. Empirical evidence shows that markers do not perceive AI as effective in addressing vague, long, or poorly aligned marking schemes and instead call for clearer rubrics, better exemplars, and more consistent interpretation. Markers emphasised confusion around marking criteria and rejected the notion that AI could interpret ambiguous schemes more reliably than humans. Quantitative results support this stance, with very low perceived AI effectiveness for addressing marking scheme vagueness or irrelevance.

MDMF Version 5, therefore, positions pre-marking tasks as deliberate human sense-making activities, including calibration for AI-fluent but conceptually weak writing, explicit discussion of originality and criticality in the AI era, and exemplars that prioritise intellectual depth over surface polish. AI awareness is reframed as a calibration topic rather than a detection or automation exercise.

Table 12 presents the operational design of calibration in MDMF v5, while Section "Fairness, bias, and AI-mediated distortions of judgment" of the Results provides empirical evidence showing how misalignment, AI-mediated fluency, and reader–supervisor asymmetries currently necessitate human calibration practices."

**Table 12.** Core component 3 (Revised). Pre-Marking Tasks—Operational Clarifications in MDMF v5.

| Aspect | MDMF v5 Update |
|---|---|
| Calibration | Repositioned as the primary human mechanism for managing AI-related uncertainty, including misalignment caused by AI-polished but conceptually weak work |
| Marking schemes | Emphasis on reducing ambiguity and interpretive drift rather than outsourcing interpretation to AI tools |
| Exemplars | Prioritise critical engagement and evidential depth over linguistic polish, including exemplars of AI-fluent but weak cases |
| AI awareness | Treated as a calibration and discussion topic for markers, not a detection or automation task |

#### 6.2.4. Core Component 4 (Clarified): Assigning Marking Based on Marker Characteristics

MDMF v5 sharpens this component by highlighting information asymmetry between supervisors and readers as a key fairness risk exacerbated by AI. Empirical findings show that readers often lack insight into the project process, particularly when AI-assisted fluency masks weak methodological or developmental engagement. This can lead to moderation conflict and misaligned judgments. Markers also reported challenges related to topic diversity and lack of subject familiarity, occasionally using AI for contextual understanding but remaining cautious about its interpretive limits. Quantitative data reinforce the prominence of workload and expertise mismatch issues, suggesting that allocation decisions remain central to equitable marking.

MDMF Version 5, therefore, emphasises structured information sharing, clearer supervisor summaries, and allocation practices that explicitly consider project complexity and expertise alignment. AI expertise is treated as contextual literacy rather than a substitute for disciplinary judgment.

Table 13 summarises refinements to Core Component 4 in MDMF Version 5. The updated framework clarifies how AI-mediated fluency intensifies existing asymmetries between supervisors and readers, reinforcing the need for structured context sharing, careful expertise matching, and informed allocation practices. AI literacy is positioned as contextual awareness rather than a source of evaluative authority, reaffirming the primacy of disciplinary judgment in marker assignment.

**Table 13.** Core component 4 (Revised). Assigning Marking Based on Marker Characteristics—Clarifications Introduced in MDMF v5.

| Dimension | MDMF v5 Refinement |
|---|---|
| Supervisor–reader asymmetry | Explicitly recognised as a fairness risk intensified by AI-polished writing |
| Reader insight | Greater emphasis on structured supervisor summaries and project-context sharing |
| Expertise matching | Reaffirmed as essential in managing topic diversity and AI-mediated surface quality |
| AI literacy | Defined as contextual understanding, not evaluative authority |

#### 6.2.5. Core Component 5 (Expanded): Marking Process

MDMF v 5 deepens the marking process component by explicitly integrating marker well-being, cognitive load, and emotional fatigue as structural considerations rather than incidental effects. Staff consistently reported workload pressure, time constraints, and frustration with repetitive, AI-fluent language patterns that reduce engagement and increase emotional strain. Quantitative data confirm that high marking load and time pressure are among the most salient challenges, while AI is not perceived as an effective solution to these issues. Although some markers noted that AI-polished writing may be easier

to read, this was offset by concerns about verbosity, loss of authenticity, and reduced meaningful differentiation.

MDMF Version 5, hence, reframes marking culture as requiring intentional process redesign, including staged assessment, improved moderation protocols, and institutional recognition of emotional labour. AI awareness is integrated into moderation discussions as a contextual factor rather than a basis for punitive or automated judgment.

Table 14 outlines the expanded focus of Core Component 5 in MDMF Version 5. The revisions position marking processes as socio-emotional and cognitive practices, explicitly integrating marker well-being, workload realism, and discussion-based moderation as central conditions for fair assessment in AI-mediated contexts, rather than treating them as secondary or individual concerns.

**Table 14.** Core component 5 (Revised). Marking Process—Expanded Focus in MDMF v5.

| Element | MDMF v5 Enhancement |
|---|---|
| Marker well-being | Integrated as a structural condition, not an ancillary concern |
| Cognitive load | Recognised impact of repetitive AI-generated language on engagement and fatigue |
| Moderation | Emphasises discussion-based resolution over algorithmic or suspicion-driven approaches |
| Workload realism | Explicit acknowledgement that AI does not resolve time and volume pressures |

#### 6.2.6. Core Component 6 (Concretised): Technology, Infrastructure, and Systems Design

In MDMF v5, the Technology component moves beyond abstract concerns about system proliferation to explicitly address submission and marking infrastructure quality as a determinant of fairness and efficiency. Empirical data highlight practical frustrations related to fragmented submission systems, unclear project metadata, burdensome supplementary materials, including long videos, and poorly integrated rubrics. Markers identified these infrastructural issues as more immediately impactful on marking quality than AI tools themselves. Quantitative findings support this, showing low perceived effectiveness of AI for addressing anonymity, contribution clarity, or process coordination.

MDMF Version 5 thus reframes technology as an enabling environment that should reduce friction, support transparency, and facilitate human judgment. AI features are positioned as optional, peripheral supports rather than core evaluative tools, with strong emphasis on data protection, traceability, and staff control.

Table 15 presents the reframing of Core Component 6 in MDMF Version 5. The updates reposition technology as an enabling infrastructure that supports transparency, efficiency, and human judgment, emphasising system design, ethical data practices, and traceability over the expansion of AI-driven assessment functions.

**Table 15.** Core component 6 (Revised). Technology and Infrastructure—Reframing in MDMF v5.

| Area | MDMF v5 Update |
|---|---|
| System design | Focus on reducing friction rather than adding AI functionality |
| Submission infrastructure | Identified as a primary determinant of marking efficiency and fairness |
| AI tooling | Positioned as optional, peripheral support rather than a core marking technology |
| Data ethics | Reaffirms prohibition of training AI on student work without consent |
| Transparency | Infrastructure must support traceability of decisions, not automated judgment |

The Master's Dissertation Marking Framework (MDMF) Version 5, the final version presented in this paper, is refined using mixed-methods evidence collected from academic staff who mark MSc dissertations in the post–generative-AI context. MDMF v5 consolidates six interdependent core components retained from Version 4 (see Figure 5), which collectively address fairness, equity, professional judgment, and institutional responsibility in dissertation assessment. While the structural architecture shown in Figure 5 remains unchanged, MDMF v5 re-weights and operationalises each component to reflect empirically grounded shifts in marking practice following the widespread adoption of generative AI tools. The framework foregrounds ethical boundary-setting, human calibration, infrastructural design, and marker well-being as central conditions for fair and defensible assessment in AI-mediated academic environments. For this reason, no separate visual representation is introduced for MDMF v5; instead, Tables 9–14 document the empirical refinements applied to each component.

## 7. Discussion

This section synthesises findings across all phases of the study to examine how the MDMF supports fairer dissertation assessment and contributes to existing assessment practices.

### *7.1. Framework Effectiveness: How the MDMF Supports Fairer Dissertation Assessment*

The Master's Dissertation Marking Framework (MDMF) was developed to address persistent challenges in dissertation assessment, including subjectivity, inconsistency, workload pressures, and the limitations of generic marking criteria. While the framework is not intended as a predictive or algorithmic tool, its effectiveness lies in making the conditions under which fair and defensible academic judgment can occur more visible, structured, and supportable. The following sections explain how the MDMF supports fairer assessment across the empirical phases of the study and how it improves existing assessment practices. These contributions reflect the iterative, design-based development of the framework, in which data insights from successive phases informed the refinement of both its structure and underlying assumptions.

#### 7.1.1. Enhancing Consistency Through Structured Processes and Calibration

A central contribution of the MDMF is its emphasis on structured pre-marking tasks, calibration practices, and shared marking culture as mechanisms for improving consistency. Across all phases, participants reported variability in how marking criteria were interpreted, particularly when marking across diverse topics or unfamiliar domains. The framework addresses this by (a) introducing explicit pre-marking calibration activities, (b) encouraging the use of discipline-relevant exemplars, and (c) promoting collective discussion of standards prior to and during marking.

Rather than assuming that consistency can be achieved through rubrics alone, MDMF recognises that consistency emerges through social and professional alignment among markers. This aligns with findings across phases that markers rely on informal calibration strategies (e.g., reading multiple dissertations before assigning final grades) to stabilise judgments. In this way, the framework improves consistency not by eliminating subjectivity, but by making interpretive processes more transparent, shared, and accountable.

#### 7.1.2. Supporting Fairness Through Contextual and Relational Judgment

The MDMF reconceptualises fairness as a contextual and negotiated practice, rather than a purely procedural outcome. While traditional mechanisms such as double marking and moderation remain important, data show that these alone do not resolve tensions related to supervisor familiarity, variation in project complexity, language differences, and differing interpretations of quality.

The framework supports fairer assessment by explicitly identifying sources of bias and inequity (e.g., language bias, fluency bias, subject expertise), embedding these considerations into marking processes and training, and encouraging reflective and dialogic moderation practices. Importantly, MDMF shifts the focus from attempting to eliminate bias entirely to recognising, managing, and negotiating bias within professional judgment. This enables a more realistic and defensible approach to fairness in complex, high-stakes assessment contexts.

#### 7.1.3. Improving Transparency and Accountability in Decision-Making

A key challenge identified across phases was the lack of transparency in how marks are derived, particularly when markers rely on tacit standards or intuitive judgment. This can lead to inconsistencies, disputes, and reduced trust in assessment outcomes. The MDMF improves transparency by (a) structuring marking processes into clearly defined components (e.g., pre-marking, allocation, moderation), (b) encouraging documentation of moderation decisions, and (c) integrating technology-supported traceability (e.g., logging discussions, recording rationale for decisions). In the post-generative-AI context, transparency becomes even more critical, particularly in cases where AI use is suspected or declared. The framework introduces processes for handling ambiguity in authorship, documenting decisions related to AI use, and ensuring that assessment decisions remain auditable and defensible. This resonates with recent work showing that transparency in assessment is a dynamic and contextually situated practice rather than a static property of documentation, requiring interaction, dialogue, and supporting infrastructures to be realised in practice (Gonsalves & Lin, 2025).

#### 7.1.4. Addressing AI-Mediated Fairness Risks

One of the most significant contributions of the MDMF is its ability to respond to new fairness challenges introduced by generative AI, which are not adequately addressed by existing assessment practices. Findings across the study indicate that AI does not reduce traditional biases such as supervisor or reader bias, introduces new risks such as fluency bias and grade compression, and complicates judgments of originality, authorship, and student contribution.

The framework addresses these challenges by incorporating AI-aware calibration practices, introducing ethical boundaries and policy clarity as a core component, and reframing fairness risks as systemic and contextual, rather than purely individual issues. Rather than positioning AI as a solution to assessment challenges, the MDMF integrates AI awareness into human-led processes, ensuring that technological change does not undermine fairness.

#### 7.1.5. Supporting Marker Well-Being and Sustainable Assessment Practices

An important but often overlooked factor in fair assessment is the well-being and cognitive capacity of markers. Across multiple phases, participants reported high workload, time pressure, fatigue, and emotional strain, particularly when marking repetitive or AI-generated text. The MDMF explicitly integrates marker well-being and emotional engagement as structural components of the assessment process. It recognises that fatigue and cognitive overload can affect the quality of judgment, and that sustainable assessment practices are necessary to maintain fairness over time. By incorporating considerations such as workload distribution, emotional support, and reflective spaces for discussion, the framework contributes to more sustainable and equitable marking practices.

#### 7.1.6. Providing a Coherent, Adaptable Structure for Complex Assessment Contexts

Finally, the effectiveness of the MDMF lies in its ability to provide a coherent yet adaptable structure that integrates multiple dimensions of assessment, including marking criteria and rubrics, assessor roles and expertise, moderation processes, institutional policy, and technological infrastructure. Unlike isolated interventions (e.g., introducing a new rubric or moderation procedure), the framework offers a holistic model that recognises the interdependence of these elements. At the same time, it is not prescriptive. Institutions can adapt the framework to their disciplinary contexts, assessment cultures, and regulatory requirements. This balance between structure and flexibility enables the MDMF to function as both a conceptual model and a practical tool, supporting fairer assessment across diverse settings.

### *7.2. Framework Contribution: Positioning the Master's Dissertation Marking Framework (MDMF) Within Assessment Theory*

This study contributes to assessment theory by introducing the Master's Dissertation Marking Framework (MDMF), a human-centred, empirically grounded framework that addresses the epistemic, ethical, and professional challenges posed by generative AI in high-stakes dissertation assessment. While assessment research has long recognised dissertations as complex, judgment-based artefacts resistant to full standardisation (D. R. Sadler, 1989; Boud & Falchikov, 2007), MDMF demonstrates that generative AI does not resolve subjectivity; rather, it reshapes the conditions under which professional judgment is exercised.

#### 7.2.1. Reasserting Professional Judgment in High-Stakes Assessment

A central theoretical contribution of MDMF is its explicit reassertion of professional judgment as the foundation of valid dissertation assessment. D. R. Sadler (1989) argued that evaluative judgment relies on assessors' capacity to interpret quality in relation to disciplinary standards rather than mechanical rule application. Similarly, Boud and Falchikov (2007) emphasised that higher-order academic work demands expert interpretation grounded in disciplinary knowledge and experience.

MDMF extends this tradition by showing that generative AI introduces new interpretive risks, such as fluency bias, masked superficiality, and uncertainty over authorship, that cannot be meaningfully resolved through automation. Rather than positioning AI as a corrective to human bias, the framework conceptualises AI as a contextual disruptor that increases the importance of assessor expertise, calibration, and reflective judgment. In doing so, MDMF challenges techno-solutionist approaches to assessment and reinforces the theoretical position that valid assessment of complex work remains irreducibly human.

#### 7.2.2. From Procedural Fairness to Contextual Equity

Assessment research has traditionally framed fairness through procedural mechanisms such as anonymity, double marking, and moderation, emphasising consistency and standardisation as safeguards against bias (Crisp, 2013; Gipps, 1994). While these remain important, MDMF demonstrates that generative AI complicates procedural fairness without necessarily improving equity. Empirical findings show that markers do not perceive AI tools as effective in addressing core fairness challenges such as supervisor bias, reader bias, or language bias, despite moderate perceived usefulness for workload-related tasks.

This aligns with socio-cultural perspectives on assessment, which argue that fairness is not merely procedural consistency, but a contextual and negotiated practice shaped by disciplinary norms, professional dialogue, and shared understanding of standards (Crisp, 2013; O'donovan et al., 2004). MDMF reconceptualises fairness as a contextual

and negotiated practice achieved through marking culture, calibration, and institutional support, rather than through algorithmic uniformity.

#### 7.2.3. Assessment as a Sociotechnical System

MDMF advances assessment theory by explicitly conceptualising dissertation marking as a sociotechnical system, in which fairness and validity are shaped by interactions between people, policies, technologies, and organisational structures. This builds on critical scholarship that cautions against viewing educational technologies as neutral tools, instead highlighting their role in shaping values, behaviours, and power relations (Williamson & Eynon, 2020). This is consistent with recent work that positions assessment as socially situated and shows that evaluative judgement emerges through contextual practices rather than from formal assessment designs alone (Fischer et al., 2024).

Rather than advocating AI-driven grading or automated decision-making, MDMF positions technology as an enabling or constraining infrastructure that must support transparency, traceability, and human deliberation. This aligns with Mittelstadt's (2019) argument that ethical AI governance requires clear accountability structures and cannot rely solely on technical fixes. By embedding ethical boundaries and institutional responsibility into the framework, MDMF offers a theoretically grounded alternative to efficiency-driven models of AI adoption in assessment.

#### 7.2.4. Emotional Labour and Marker Well-Being as Conditions of Fair Assessment

A distinctive theoretical contribution of MDMF is the explicit integration of marker emotional labour and well-being into the assessment framework. While workload and stress are acknowledged in higher education research, they are rarely theorised as conditions that directly affect assessment quality and fairness. Emerging scholarship on academic labour highlights how emotional exhaustion, moral discomfort, and cognitive overload shape professional practice (Kinman & Wray, 2018).

MDMF extends assessment theory by demonstrating that AI-mediated marking introduces new affective pressures, including fatigue from repetitive AI-generated language and anxiety surrounding suspected but unverifiable AI use. By positioning emotional engagement and well-being as structural components rather than incidental concerns, the framework aligns with calls for more sustainable and humane models of academic work in higher education.

#### 7.2.5. Boundary-Setting in the Post-Generative-AI Era

Finally, MDMF contributes to assessment theory by shifting the focus from regulating how AI might be used in marking to defining what must remain human in high-stakes academic judgment. Rather than attempting to optimise AI integration, the framework establishes principled boundaries around authorship evaluation, ethical responsibility, and academic standards. This stance resonates with contemporary critiques of automated assessment, which warn that delegating evaluative judgment to AI risks undermining both trust and educational values (Bearman et al., 2024). In this sense, MDMF offers a boundary-setting framework that preserves the epistemic and ethical foundations of dissertation assessment while acknowledging the realities of AI-mediated academic work. This aligns with emerging scholarship arguing that GenAI may support educational processes without reproducing the recognitive, relational, and trust-based dimensions of human pedagogical judgment (Corbin et al., 2025). In dissertation marking, these limitations are especially significant because evaluative decisions are high-stakes, disciplinary, and ethically consequential. Related evidence suggests that AI-generated feedback is often perceived as less credible and less genuine once its source is disclosed (Nazaretsky et al., 2026), and that the

educational value of GenAI depends heavily on human literacies and critical oversight rather than arising automatically from the technology itself (Zhan et al., 2025).

### *7.3. Implications for Practice*

As a framework developed to address dissertation marking in practice, the MDMF has several practical implications for institutions seeking to support fairer dissertation assessment. First, the framework highlights that fairness cannot be achieved through marking schemes alone, but requires structured calibration, shared interpretation of standards, and ongoing dialogue among markers. Institutions should therefore prioritise pre-marking calibration and discussion-based moderation practices alongside the use of rubrics. Second, the findings underline the importance of clear institutional guidance on generative AI. Ambiguity around acceptable AI use introduces inconsistency and uncertainty in marking. The MDMF supports the development of transparent policies and aligned marking practices that clarify expectations for both students and staff. Finally, the framework demonstrates that fair assessment depends on broader assessment conditions, including workload, marking processes, and infrastructure. Supporting marker alignment, reducing administrative friction, and recognising the demands of dissertation marking are essential for maintaining consistency and fairness in practice.

### *7.4. Limitations and Future Research*

There are several limitations. First, the empirical work was conducted within a single institutional context, which may limit transferability to other disciplines or settings. Second, the study draws on small, purposive samples, particularly in the interview phases. While consistent with design-based research, this limits the generalisability of findings but supports their relevance for framework development. Third, while the MDMF has been iteratively refined across multiple phases, it has not yet been fully implemented and evaluated as a complete intervention. Future research should examine its application in practice, including its impact on marking consistency and perceptions of fairness. Finally, as generative AI continues to evolve, the challenges identified in this study may shift. Further research is needed to explore how assessment practices adapt over time and to incorporate student perspectives on fairness, authorship, and AI use.

## 8. Conclusions

This paper presented the Master's Dissertation Marking Framework (MDMF), an empirically grounded framework refined through the synthesis of literature and pre- and post-generative-AI empirical studies. Building on earlier iterations, the framework responds to enduring challenges in master's dissertation assessment—such as bias, subjectivity, workload pressure, and reliability—while consolidating evidence on the ethical, epistemic, and practical tensions introduced by generative AI. MDMF comprises six interrelated core components: (1) ethical boundaries and AI policy clarity, (2) identification of fairness and equity issues, (3) pre-marking tasks, (4) marker allocation based on marker characteristics, (5) marking processes and culture, and (6) technology as both enabler and disruptor.

Instead of positioning technology as a solution to assessment subjectivity, MDMF demonstrates that fairness in dissertation marking remains fundamentally dependent on professional judgment, institutional clarity, and collective calibration. While technology can support transparency, traceability, and process efficiency, its role must remain bounded by ethical governance and clear institutional expectations about which evaluative judgments must remain human. Unregulated or poorly integrated AI risks amplifying inequities, obscuring authorship, and undermining trust in academic judgment.

Importantly, the framework is not intended as a prescriptive checklist. Instead, MDMF functions as a conceptual and practical scaffold that institutions can adapt to their disciplinary contexts, assessment cultures, and regulatory environments. By foregrounding ethical boundaries, marker well-being, and the sociotechnical conditions of assessment, this framework contributes to ongoing debates on how fairness and validity can be sustained in high-stakes assessment in the age of generative AI.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and approved by the School of Education Research Ethics Committee (protocol code 402210101, December 2021) and by the College of Science and Engineering Ethics Committee (protocol codes: 300220170, 13 June 2023; 300230028, 18 October 2023; 300240064, 10 January 2025).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to the corresponding author.

**Acknowledgments:** The author thanks the academic staff at the university who generously shared their time and experiences by participating in this research over multiple years.

**Conflicts of Interest:** The author declares no conflicts of interest.

## References

Afreen, J., Mohaghegh, M., & Doborjeh, M. (2025). Systematic literature review on bias mitigation in generative AI. *AI Ethics*, *5*, 4789–4841. [CrossRef]

An, Y., Yu, J. H., & James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education*, *22*(1), 1–23. [CrossRef]

Armstrong, M., Dopp, C., & Welsh, J. (2018). Design-based research. In R. Kimmons (Ed.), *The students' guide to learning design and research*. EdTech Books. Available online: https://edtechbooks.org/studentguide/design-based_research (accessed on 5 January 2024).

Ashworth, M., Bloxham, S., & Pearce, L. (2010). Examining the tension between academic standards and inclusion for disabled students: The impact on marking of individual academics' frameworks for assessment. *Studies in Higher Education*, *35*(2), 209–223. [CrossRef]

Bearman, M., & Ajjawi, R. (2018). From "seeing through" to "seeing with": Assessment criteria and the myths of transparency. *Frontiers in Education*, *3*, 96. [CrossRef]

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *49*(6), 893–905. [CrossRef]

Bikanga Ada, M. (2014, July 7–9). *A case study at a Scottish university*. 6th International Conference on Education and New Learning Technologies, Barcelona, Spain.

Bikanga Ada, M. (2018). Using design-based research to develop a Mobile Learning Framework for Assessment Feedback. *Research and Practice in Technology Enhanced Learning*, *13*(3), 1–22. [CrossRef]

Bikanga Ada, M. (2023, December 4–7). *I betrayed my ethical principles: Investigating master's dissertation marking practices in CS*. 2022 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), Hung Hom, Hong Kong.

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209–220. [CrossRef]

Bloxham, S., & Boyd, P. (2012). Accountability in grading student work: Securing academic standards in a twenty-first century quality assurance context. *British Educational Research Journal*, *38*(4), 615–634. [CrossRef]

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36*(6), 655–670. [CrossRef]

Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education*, *41*(4), 638–653. [CrossRef]

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

Bourke, S., & Holbrook, A. P. (2013). Examining PhD and research masters theses. *Assessment & Evaluation in Higher Education*, *38*(4), 407–416. [CrossRef]

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. [CrossRef]

Burger, R. (2017). Student perceptions of the fairness of grading procedures: A multilevel investigation of the role of the academic environment. *Higher Education*, *74*, 301–320. [CrossRef]

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. [CrossRef]

Chai, F., Ma, J., Wang, Y., Zhu, J., & Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology*, *15*, 1221177. [CrossRef]

Chakraborty, S., Dann, C., Mandal, A., Dann, B., Paul, M., & Haeez-Baig, A. (2021). Effects of rubric quality on marker variation in higher education. *Studies in Educational Evaluation*, *70*, 100997. [CrossRef]

Corbin, T., Tai, J., & Flenady, G. (2025). Understanding the place and value of GenAI feedback: A recognition-based framework. *Assessment & Evaluation in Higher Education*, *50*(5), 718–731. [CrossRef]

Crisp, V. (2013). Criteria, comparison and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice*, *20*(1), 127–144. [CrossRef]

Çağlar-Özhan, Ş., Tekeli, P., & Arkün-Kocadere, S. (2025). Comparison of AI-generated and instructor feedback: No significant difference in perceived feedback quality and neither on performance. *Journal of Computer Assisted Learning*, *41*(5), e70134. [CrossRef]

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. [CrossRef]

Elliot, E., Pearce, K., & King, S. (2011, July 4–7). *Moderation of assessment tasks: Developing solutions to common problems* [Paper presentation]. HERDSA Conference 2011, Gold Coast, Australia.

Ernstzen, D., Leibbrandt, D., & Louw, Q. (2026). The use and influence of supervisor marking in graduate research project assessment: A scoping review. *Assessment & Evaluation in Higher Education*, 1–20. [CrossRef]

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Fischer, J., Bearman, M., Boud, D., & Tai, J. (2024). How does assessment drive learning? A focus on students' development of evaluative judgement. *Assessment & Evaluation in Higher Education*, *49*(2), 233–245. [CrossRef]

Forsyth, R., Cullen, R., Ringan, N., & Stubbs, M. (2015). Supporting the development of assessment literacy of staff through institutional process change. *London Review of Education*, *13*(3), 34–41. Available online: https://files.eric.ed.gov/fulltext/EJ1160159.pdf (accessed on 25 November 2025). [CrossRef]

Gipps, C. (1994). *Beyond testing: Towards a theory of educational assessment* (1st ed.). Routledge. [CrossRef]

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith, & J. J. Cumming (Eds.), *Educational assessment in the 21st century*. Springer. [CrossRef]

Gonsalves, C. (2024). Addressing student non-compliance in AI use declarations: Implications for academic integrity and assessment in higher education. *Assessment & Evaluation in Higher Education*, *50*(4), 592–606. [CrossRef]

Gonsalves, C., & Lin, Z. (2025). Clear in advance to whom? Exploring 'transparency' of assessment practices in UK higher education institution assessment policy. *Studies in Higher Education*, *50*(7), 1454–1470. [CrossRef]

Gordon, M. E., & Fay, C. H. (2010). The effects of grading and teaching practices on students' perceptions of grading fairness. *College Teaching*, *58*(3), 93–98. [CrossRef]

Harper, G. (2016). Creative writing you're cognitive dissonance. *New Writing*, *13*(1), 1–2. [CrossRef]

Hawe, E., Dixon, H., Murray, J., & Chandler, S. (2021). Using rubrics and exemplars to develop students' evaluative and productive knowledge and skill. *Journal of Further and Higher Education*, *45*(8), 1033–1047. [CrossRef]

Hudson, J., Bloxham, S., den Outer, B., & Price, M. (2017). Conceptual acrobatics: Talking about assessment standards in the transparency era. *Studies in Higher Education*, *42*(7), 1309–1323. [CrossRef]

Jackson, D. (2018). Challenges and strategies for assessing student workplace performance during work-integrated learning. *Assessment & Evaluation in Higher Education*, *43*(4), 555–570. [CrossRef]

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(11), e2208839120. [CrossRef]

Jönsson, A., & Prions, F. (2019). Transparency in assessment—Exploring the influence of explicit assessment criteria. *Frontiers in Education*, *3*, 119. [CrossRef]

Kinman, G., & Wray, S. (2018). Presenteeism in academic employees-occupational and individual factors. *Occupational Medicine*, *68*(1), 46–50. [CrossRef]

Man, D., Xu, Y., Chau, M. H., O'Toole, J. M., & Shunmugam, K. (2020). Assessment feedback in examiner reports on master's dissertations in translation studies. *Studies in Educational Evaluation*, *64*, 100823. [CrossRef]

Marr, L., & Forsyth, R. (2010). *Identity crisis: Working in higher education in the 21st century*. Trentham Books.

McKenney, S. E., & Reeves, T. C. (2012). *Conducting educational design research*. Routledge.

McQuade, R., Kometa, S., Brown, J., Bevitt, D., & Hall, J. (2020). Research project assessments and supervisor marking: Maintaining academic rigour through robust reconciliation processes. *Assessment & Evaluation in Higher Education*, *45*(8), 1181–1191. [CrossRef]

Mei, P., Brewis, D. N., Nwaiwu, F., Sumanathilaka, D., Alva-Manchego, F., & Demaree-Cotton, J. (2025). If ChatGPT can do it, where is my creativity? Generative AI boosts performance but diminishes experience in creative writing. *Computers in Human Behavior: Artificial Humans*, *4*, 100140. [CrossRef]

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. [CrossRef]

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2026). Who gives feedback matters: Student biases towards human and AI-generated formative feedback. *Journal of Computer Assisted Learning*, *42*(1), e70153. [CrossRef]

O'donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education*, *9*(3), 325–335. [CrossRef]

Peeters, M. J., Schmude, K. A., & Steinmiller, C. L. (2014). Inter-rater reliability and false confidence in precision: Using standard error of measurement within PharmD admissions essay rubric development. *Currents in Pharmacy Teaching and Learning*, *6*(2), 298–303. [CrossRef]

Perkins, M., & Roe, J. (2024). Decoding academic integrity policies: A corpus linguistics investigation of AI and other technological threats. *Higher Education Policy*, *37*, 633–653. [CrossRef]

Petricini, T., Zipf, S., & Wu, C. (2025). RESEARCH-AI: Communicating academic honesty: Teacher messages and student perceptions about generative AI. *Frontiers in Communication*, *10*, 1544430. [CrossRef]

Pérez-Ros, P., Chust-Hernández, P., Ibáñez-Gascó, J., & Martínez-Arnau, F. M. (2021). An undergraduate thesis training course for faculty reduces variability in student evaluations. *Nurse Education Today*, *96*, 104619. [CrossRef]

Pilcher, N. (2011). The UK postgraduate masters dissertation: An 'elusive chameleon'? *Teaching in Higher Education*, *16*(1), 29–40. [CrossRef]

Pitt, E., & Winstone, N. (2018). The impact of anonymous marking on students' perceptions of fairness, feedback and relationships with lecturers. *Assessment & Evaluation in Higher Education*, *43*(7), 1183–1193. [CrossRef]

Postmes, L., Bouwmeester, R., de Kleijn, R., & van der Schaaf, M. (2022). Supervisors' untrained postgraduate rubric use for formative and summative purposes. *Assessment & Evaluation in Higher Education*, *48*(1), 41–55. [CrossRef]

Pufpaff, L. A., Clarke, L., & Jones, R. E. (2015). The effects of rater training on inter-rater agreement. *Mid-Western Educational Researcher*, *27*(2), 117–141.

Quality Assurance Agency. (2006). *Code of practice for the assurance of academic quality and standards in higher education, Section 6: Assessment of students*. QAA. Available online: https://dera.ioe.ac.uk/9713/2/COP_AOS.pdf (accessed on 23 May 2025).

Quality Assurance Agency. (2018). *UK quality code for higher education: Advice and guidance—Assessment*. Available online: https://www.qaa.ac.uk/the-quality-code/2018/advice-and-guidance-18/assessment (accessed on 23 January 2026).

Sadler, D. (2014). The futility of attempting to codify academic achievement standards. *Higher Education*, *67*(3), 273–288. [CrossRef]

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. [CrossRef]

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179. [CrossRef]

Satchell, S., & Pratt, J. (2010). The dubiety of double marking. *Higher Education Review*, *42*(2), 59–62. Available online: https://eric.ed.gov/?id=EJ877250 (accessed on 10 January 2025).

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*, 467–481. [CrossRef]

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A review of literature on marking reliability research (report for ofqual)*. NFER. Available online: https://assets.publishing.service.gov.uk/media/5a81cc0e40f0b62302699360/0613_JoTisi_et_al-nfer-a-review-of-literature-on-marking-reliability.pdf (accessed on 15 September 2025).

Vinton, L., & Wilke, D. (2011). Leniency bias in evaluating clinical social work student interns. *Clinical Social Work Journal*, *39*(3), 288–295. [CrossRef]

Williams, L., & Kemp, S. (2019). Independent markers of master's theses show low levels of agreement. *Assessment & Evaluation in Higher Education*, *44*(5), 764–771. [CrossRef]

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology*, *45*(3), 223–235. [CrossRef]

Winstone, N. E., & Boud, D. (2022). The need to disentangle assessment and feedback in higher education. *Studies in Higher Education*, *47*(3), 656–667. [CrossRef]

Wolf, K. (2015). Leniency and halo bias in industry-based assessments of student competencies: A critical, sector-based analysis. *Higher Education Research & Development*, *34*(5), 1045–1059. [CrossRef]

Zhan, Y., Boud, D., Dawson, P., & Yan, Z. (2025). Generative artificial intelligence as an enabler of student feedback engagement: A framework. *Higher Education Research & Development*, *44*(5), 1289–1304. [CrossRef]