Engaging Teachers in Co-Designing Examinations for Secondary Schools in the Era of Large Language Models

Erik Winerö[1,*] and Marie Utterberg Modén[1]

[1] *University of Gothenburg, Department of Applied IT, Gothenburg, Sweden*

Abstract

This study investigates the role of large language models (LLMs) in co-designing examination tasks for secondary school teachers. Using end-user development (EUD) principles, we explore how teachers can design examinations that either restrict or incorporate generative AI. Through a series of workshops with 153 teachers from 7 schools, we gathered qualitative data on their current assessment practices and their interactions with AI. Our findings reveal diverse strategies for integrating AI in educational assessments, highlighting both opportunities and challenges. This research underscores the importance of professional development to enhance teachers' AI literacy and improve the congruence between their technological frames and pedagogical practices.

Keywords

Large language models, co-design, educational assessment, examinations, secondary education, AI literacy, end-user development.

1. Introduction

There is a growing interest among researchers in the field of educational technology to employ participatory design methods that allow for the empowerment of teachers through genuine engagement with a relevant problem [1]. Throughout its history, participatory design has actively engaged in processes that aim to foster new insights, skills, visions, and democratic awareness among individuals by involving them in design and technology initiatives. Early participatory design projects were guided by these political commitments and aimed to empower future users to actively participate in technological development [2].

However, Pérez-Sanagustín et al. [3] found in their literature review of research on technology in education that studies concerned with design of digital solutions often lack involvement and mutual meaningful collaboration with teachers. At present, AI and large language models (LLMs) have been described as the most prominent and controversial technologies to impact education [4]. Introducing AI in classrooms becomes challenging when 'available tools and curriculum are incompatible with values and contexts' of the people who use them [5].

Proceedings of the First International Workshop on Participatory Design & End-User Development - Building Bridges (PDEUD2024), October 2024, Uppsala, Sweden

[*] Corresponding author.

erik.winero@ait.gu.se (E. Winerö); marie.utterberg@ait.gu.se (M. Utterberg Modén)

0000-0002-3820-4063 (M. Utterberg Modén)

By incorporating end-user development (EUD) principles, we investigate how teachers navigate and strategize around the use of generative AI in their assessment practices, influenced by their own interpretations and understandings of the technology. While traditional EUD often involves control over design [6], in educational contexts, teachers typically have limited influence over the design of the technology they use. Therefore, we include the concept of EUD to encompass the ways teachers adapt their practices and create new workflows around existing technologies, particularly AI and LLMs. Thus, EUD involves teachers developing novel approaches to use existing technologies, customizing their pedagogical practices, and creating new assessment methodologies that either incorporate or exclude AI use.

2. Large language models in education

In May 2022, Sharples [7] published a blog post titled "New AI tools that can write student essays require educators to rethink teaching and assessment". The impetus for Sharples' post was the rapid advancement in the field of generative AI and Large Language Models that began to prominently emerge at the start of the decade. Notably, OpenAI's GPT-3, launched in 2020, showcased an unprecedented ability to produce text virtually indistinguishable from that written by humans. Sharples' contribution stands as an early instance where an educator underscored the impact of the emerging LLMs on traditional assessment methods. Since that time, a growing chorus of voices has resonated with similar concerns, prompting national and global educational institutions and bodies like UNESCO [8] to formulate guidelines addressing these challenges.

In educational settings, various forms of written examinations have long served as an essential method for assessment. The inherent reason being that declarative knowledge is conveyed through language, and written language has historically been intuitive to use from an assessment standpoint due to its physical manifestation. Spoken word is by its nature temporary and context-dependent [9]. This has made written examinations a practice deeply intertwined with the use of tools, ranging from traditional pen and paper to modern word processors. In recent decades, digital applications like grammar checks and spell checks have assisted students in refining their texts. Generally, these tools are perceived as means to polish the surface of a text without adversely affecting its ability to serve as a valid representation of the student's thoughts and learning. However, the recent rapid development and increased access to LLMs and systems such as ChatGPT, Google Gemini and Claude has come to challenge these notions.

In discussions about the role of technology in society, two competing perspectives often emerge. The first, rooted in technodeterminism, posits that it is the technology itself that drives change, shaping educational practices and outcomes almost independently of the users' intentions or understanding. An often-mentioned example is the expectations on transforming teaching and learning by introducing computers in schools. The second perspective emphasizes the role of users, arguing that the impact of technology is mediated by how users perceive, interpret, and integrate technology into their practices. This latter view challenges the notion of technology as an autonomous force, instead highlighting the importance of socio-cognitive factors in shaping its implementation. Our study aligns with this second perspective, focusing on how teachers' perceptions and interpretations of technology influence its integration into educational practices. We apply Orlikowski and Gash's concept of *technological frames* [10], extending it with Orlikowski’s notion of sociomateriality [11], to explore how teachers' assumptions, expectations and knowledge about generative AI and LLMs influence and affects teachers’ assessment practices.

From this perspective, teachers' technological frames become foundational to their reasoning, design process, and ultimately, the assessment methods they employ. By utilizing this theoretical assumption as a lens, we aim to elucidate the connection between teachers' assumptions, expectations, and knowledge, and how these shape their pedagogical practices. Specifically, we examine how teachers' technological frames regarding generative AI and LLMs influence their development of summative assessment methods. Ultimately, our goal is to contribute to a deeper understanding of the complex interplay between pedagogy and technology. Through this, we hope to empower teachers to develop assessment methods that both incorporate and exclude AI in a conscious, well-reasoned, and sustainable manner.

3. Method

The study employed a qualitative approach, conducting workshops with teachers where they were tasked with designing two types of assessment tasks: one that restricts AI use and one that integrates AI. Data were collected through observation and analysis of the tasks during 11 workshops and a total of 153 teachers from 7 schools. The first workshop took place at the beginning of March, and the last one in early June 2023.

Our research was designed to serve a dual purpose: to collect data relevant to our study and to provide professional development for the teachers involved. This approach was in response to the large number of inquiries we received from schools regarding professional development in generative AI during the winter and spring of 2023.

While this study focuses on teachers in Sweden, the sample size provides a robust basis for qualitative analysis. The participants represent a diverse group of educators across various subjects and experience levels, enhancing the study's internal validity. Moreover, the challenges and opportunities presented by generative AI in education are largely universal, transcending national boundaries. The rapid global adoption of AI technologies in education suggests that many of the insights gained from this study may be applicable to international contexts. However, we acknowledge that cultural, policy, and infrastructural differences may influence the specific ways in which AI is integrated into educational practices across different countries. Future research could explore these potential variations in more depth.

Each workshop commenced with an overview of generative AI and LLMs, aimed at equipping participants with a common understanding without swaying their perspectives. We deliberately avoided in-depth discussions about how LLMs work, such as their foundation in statistical models or the uniqueness and unpredictability of their generated texts. Our introduction was limited to a brief historical overview and a demonstration of ChatGPT, ensuring it was a primer rather than a detailed lecture. The participating teachers were then individually asked to document their most recent and frequently used assessment practices. Following this, teachers were grouped randomly into small groups to share and discuss their summative assessment practices in relation to recent development in LLMs. After this, the teachers were randomly grouped into pairs or trios with the task of designing two separate variants of examinations. One variant was to be designed in such a manner that generative AI could not be used or would not be beneficial for students to use. The other variant was to incorporate generative AI as a central component of the examination. The teachers were given about 30 minutes to design these examinations, after which they presented them to the rest of the group. These presentations were recorded and transcribed, serving as the main data for this study.

The teachers were also given an anonymous survey where they were asked to provide information about the subjects they taught, whether they held a teaching certification, how long they had been teaching, and whether they had received any specific professional development in the field of AI, and to what extent they potentially had explored generative AI independently outside of their formal work responsibilities.

It is important to highlight that we did not categorize teachers based on the subjects they taught. This decision was driven by our aim to foster a broader discourse and given that teachers in Sweden frequently work in interdisciplinary teams (as was the case with all participating teachers in this study), we chose to maintain the discussions within the study as subject transcendent.

4. Results and findings

Out of 152 teachers, 137 responded to the survey, which represents a response rate of approximately 90%. Among the respondents, 89 teachers indicated they taught at the middle school level, while 48 reported teaching at the high school level. These educators taught across a diverse range of subjects, with a notable emphasis on Humanities and Social Sciences (HUMSS) and Science, Technology, Engineering, and Mathematics (STEM), in addition to languages like Swedish and English, arts, and physical education. Their experience in the field varied significantly, averaging nearly 18 years, with a span from 1 to 35 years. Of the middle school teachers, 73 percent held teaching certifications, compared to the national figure of 71.5 percent (Swedish National Agency for Education, 2024). For high school teachers in the study, the certification rate was 83 percent, against a national average of 84.2 percent. Regarding the teachers' previous experience of generative AI and LLMs the answers from the survey were categorized into three categories: *none*, *some*, and *extensive*, with none of the responses being considered as meeting the criteria for the latter (fig. 1). Responses categorized as 'some' typically referred to professional development consisting of a short lecture or information session about ChatGPT, and when it came to non-professional experience, it mostly involved teachers having tested and explored generative AI services on their own on one or a few occasions (fig. 2).

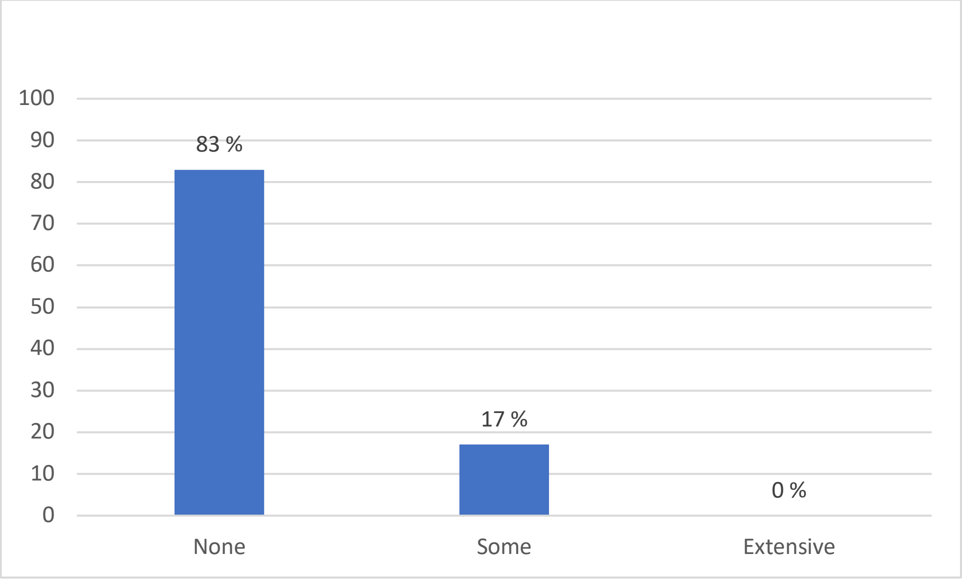

Figure 1:. Previous Professional Development in Generative AI such as ChatGPT

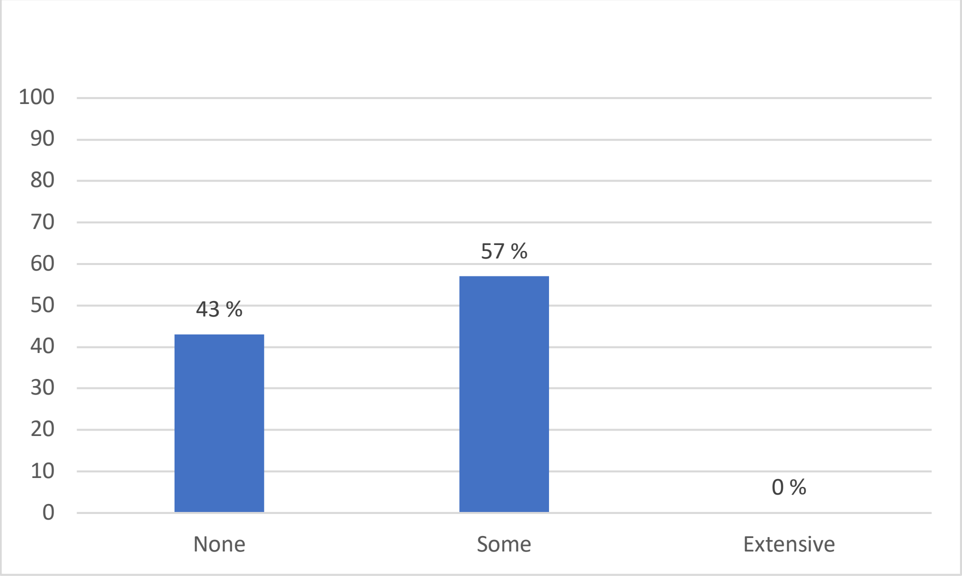

Figure 2: Prior Non-Professional Experience with Generative AI such as ChatGPT

In the task of designing an assessment that would exclude AI an overwhelming majority (93 %) of the teachers chose to employ various forms of traditional assessment situations, where the environment and tools were controlled and restricted to limit students' access to generative AI (fig. 3). This was typically achieved through a) traditional in-class exams by returning to paper-and-pencil exams conducted in the classroom, b) digital assessment system by utilizing digital platforms that limit internet access, thereby preventing the use of AI tools, and c) oral examinations by conducting assessments orally to ensure that AI cannot be used during the evaluation process. Just a smaller portion of the teachers (7%) chose to design tasks they perceived to be inherently designed in such a way that students were not justified in using AI. Examples of such tasks included assignments where students were asked to write more personal texts, such as referring to their own experiences, or to write analyses of texts provided only at the time of the exam (in these and other described assessment formats, it could be argued that the teachers underestimated the capabilities of generative AI, which is why we refer to these as (perceived) AI-resistant tasks.

The overwhelming preference (93%) for traditional, controlled assessment environments to limit AI use reveals a significant trend in teachers' approaches to excluding AI from assessments. This reliance on established methods suggests that teachers may feel more confident in their ability to control the assessment environment rather than in designing inherently 'AI-proof' tasks.

The small portion (7%) of teachers who attempted to design tasks they perceived as inherently resistant to AI use presents an intriguing area for further exploration. These attempts, which included assignments focusing on personal experiences or real-time analysis of provided texts, highlight the creative approaches some educators are taking to address the challenges posed by AI. However, the limited adoption of such strategies also underscores the difficulty in designing truly 'AI-resistant' tasks in an era of rapidly advancing language models.

This stark contrast between traditional control methods and attempts at AI-resistant task design provides valuable insights into teachers' current comfort levels and perceived capabilities in navigating the AI landscape in education. It suggests a need for professional development not only in AI integration but also in designing authentic assessments that maintain their validity in an AI-rich environment without necessarily reverting to traditional, controlled testing situations.

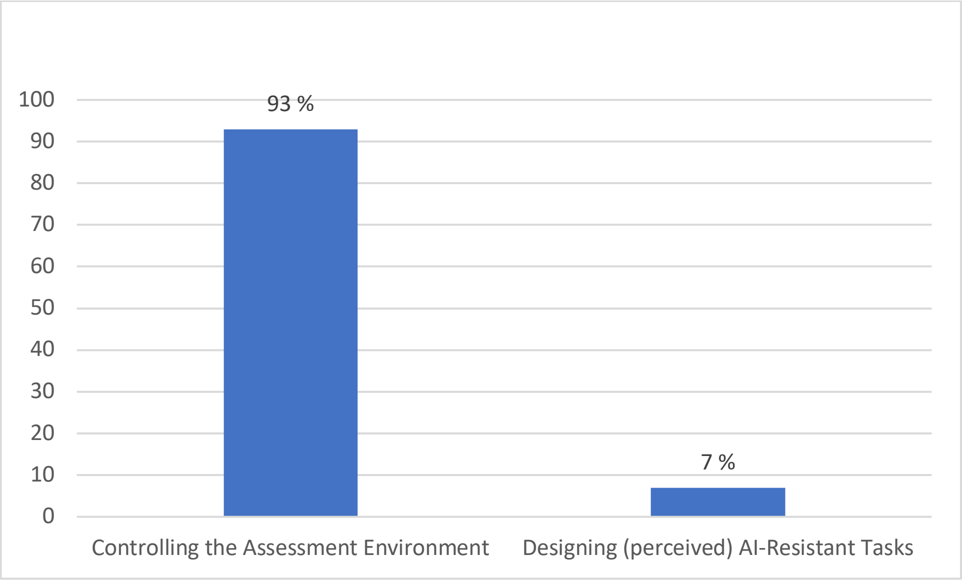

Figure 3: Distribution of Strategies Used by Teachers to Limit AI Use in Assessments

When it came to the task of designing assessments where AI would be integrated, a majority (86%) chose various approaches where students would analyze AI-generated material, such as analyzing (and in some cases even detecting!) AI-generated texts. A smaller portion (14%) chose tasks where students would in various ways build upon AI-generated material. This could involve skills allowing AI to generate basic material so that students could focus on more complex aspects of the task. For example, AI might generate the foundations of a song or a piece of writing, which students then refine and develop further. Using AI for language practice: Employing AI bots for conversational practice to enhance language skills. Fairer oral assessments: Generating scripts with AI for oral presentations, thereby assessing only the delivery and not the scriptwriting. This method aimed to create a more equitable assessment environment.

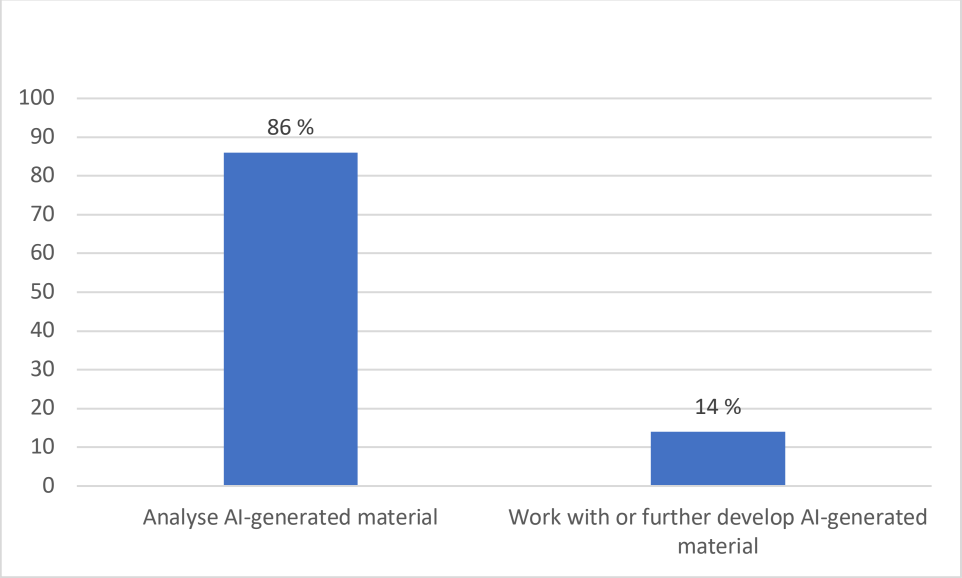

Figure 4: Distribution of Strategies Used by Teachers incorporate student use of AI in Assessments

5. Discussion

Analysis of the teachers' presentations revealed two primary perspectives on how AI can be incorporated into assessment practices. These perspectives can be illustrated through a Venn diagram (Fig. 5). The dominant approach among teachers was to view AI as a learning objective. In this context, teachers designed tasks where students were required to reflect on and analyse AI-generated content. A less prominent, yet significant perspective was the view of AI as a tool. Here, teachers provided fewer concrete examples, potentially indicating limited personal experience in working with AI. One of the few examples mentioned was allowing students to use AI to generate scripts for oral presentations, which would enable an assessment focused on delivery rather than script writing.

This discrepancy between the two perspectives can be understood through the concept of 'technological frames' [10], which highlights how teachers' own assumptions, expectations, and knowledge about AI influence their conceptions of its role in teaching and assessment. The Venn diagram serves as a visual representation of these two dimensions - AI as a learning objective and AI as a tool - and their potential overlap in assessment. It is worth noting that this conceptualization applies only to the tasks where teachers were asked to include AI, not to those where AI was to be excluded.

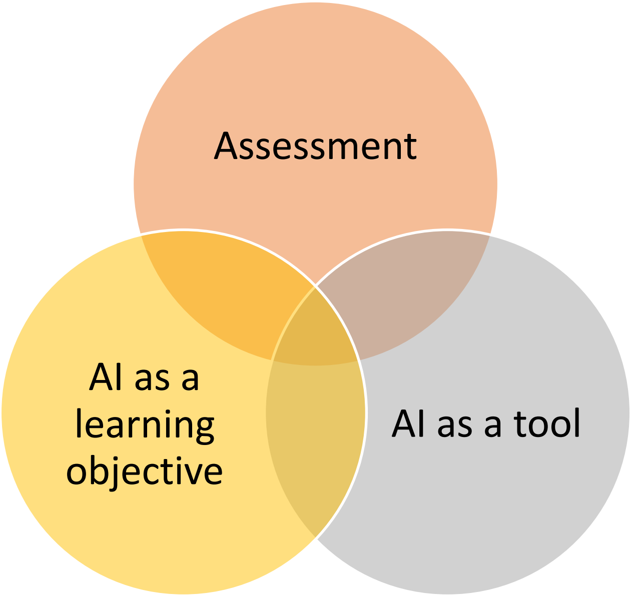

Figure 5: Venn diagram of three perspectives on how LLMs affect assessment.

This Venn diagram illustrates the complex interplay between Assessment, AI as a learning objective, and AI as a tool in the context of our study. Each circle represents a key area:

Assessment (Red circle): Traditional and evolving methods of evaluating student learning and performance.

AI as a learning objective (Purple circle): Teaching students about AI, including its capabilities, limitations, and societal impacts.

AI as a tool (Green circle): Using AI technologies to support or enhance learning and teaching processes.

The intersections of these circles represent areas where teachers are engaging in complex forms of end-user development (EUD):

Assessment + AI as a learning objective: Here, teachers design assessments that test students' understanding of AI concepts, such as the tasks where students analyse AI-generated texts.

Assessment + AI as a tool: This intersection involves using AI to support assessment processes, as seen in the tasks where AI generates basic material for students to build upon.

AI as a learning objective + AI as a tool: This area combines teaching about AI with practical use of AI tools, reflected in tasks where students both use and critically analyse AI.

The central intersection represents the most integrated approach, where AI is simultaneously a subject of study, a tool for learning, and part of the assessment process.

While the Venn diagram (Fig. 5) serves as a useful heuristic tool for visualizing the different perspectives on AI integration in assessment practices, it is important to acknowledge its limitations. As with any model, it necessarily simplifies the complex reality of teachers' attitudes and practices. The clear-cut categories and intersections may not fully capture the nuanced and sometimes conflicting views that individual teachers may hold. Despite these limitations, the Venn diagram provides insights by offering a clear, visual representation of the main themes that emerged from our analysis. It helps to conceptualize the different ways teachers approach AI in assessment and highlights potential areas of integration. Future research could build upon this model, perhaps developing more sophisticated representations that capture additional dimensions of teachers' perspectives and practices.

Having acknowledged both the utility and limitations of the Venn diagram as an analytical tool, we can now delve deeper into interpreting the patterns it helps us visualize. One particularly striking observation is the teachers' predominant focus on AI as a learning objective. We interpret this focus, for both teachers and students, as indicative of a broader phenomenon. Specifically, we see it as a sign that teachers themselves are not yet sufficiently knowledgeable or comfortable with using AI in their practice. The tasks they assign reflect their own knowledge and limitations. In other words, the assignments and the skills they assess are influenced by their own understanding, which argues against a technodeterministic perspective and supports a socio-cognitive perspective.

While our findings suggest a connection between teachers' limited experience with AI and their tendency to focus on AI as a learning objective rather than a tool, it's important to recognize that this relationship is likely more complex than a simple cause-and-effect scenario. Various factors could influence teachers' approaches, including but not limited to their prior technological experiences, pedagogical beliefs, institutional policies, and the specific subject areas they teach. Moreover, the correlation we observe between limited AI experience and a focus on AI as a learning objective might be bidirectional. Teachers with less experience might naturally gravitate towards teaching about AI rather than with it, but equally, a curriculum emphasis on AI literacy could lead to teachers spending more time learning about AI than experimenting with it as a tool.

Future research could benefit from a more granular analysis of these factors, possibly employing mixed methods to quantify the strength of various influences on teachers' AI integration strategies. This could help disambiguate correlation from causation and provide a more comprehensive understanding of how teachers' technological frames evolve in relation to their practical experiences with AI.

Given this complex interplay of factors influencing teachers' approaches to AI integration, it's crucial to consider how these observations align with broader theoretical frameworks. Particularly relevant is the concept of End-User Development (EUD) in educational contexts. This interpretation, acknowledging both the predominant focus on AI as a learning objective and the multifaceted influences shaping teachers' practices, aligns with EUD principles. In this context, teachers are developing new understandings and practices around AI rather than modifying the technology itself. The process of designing AI-integrated and AI-excluded assessments can be seen as a form of practice-oriented customization, where teachers are essentially creating new 'programs' of practice. This form of EUD is particularly relevant in educational contexts where direct technological modification is often not feasible.

The collaborative nature of the workshops also highlights the potential for community-driven EUD in education. By sharing and refining their approaches to AI in assessment, teachers engaged in a collective form of end-user development, creating shared knowledge and methodologies that can be adapted to various educational contexts.

The Venn diagram helps us visualize how teachers are navigating between different aspects of AI integration in education. As they design assessments, they're moving between these different areas, making decisions about how to balance assessment needs, AI education, and AI integration, where teachers are shifting from the ‘fast thinking’ operations typically provided by generative AI to the ‘slow thinking’ aspect of EUD, which involves critically verifying and going beyond the information presented [6]. This navigation process itself is a form of EUD, as teachers are developing new practices and understandings in a complex, evolving educational landscape.

Moreover, the diagram illustrates how teachers' technological frames (their understanding and perceptions of AI) influence their EUD activities across these areas. Teachers with different levels of AI literacy or different views on AI's role in education might focus their EUD efforts in different areas of the diagram. For instance, teachers who are more comfortable with AI might design assessments that fall in the central intersection, integrating all three aspects, while those less familiar with AI might focus more on the "Assessment" circle or the intersection of "Assessment" and "AI as a learning objective".

It is important to note that this study represents a snapshot of teachers' technological frames and practices during a specific period (March to June 2023). Given the rapid evolution of AI technologies, it is likely that these frames have since evolved. However, this temporal specificity does not diminish the study's value. Rather, it underscores the dynamic nature of technological frames and their impact on educational practices. The key insight lies not in the specific content of the frames at that time, but in demonstrating how these frames shape teachers' approaches to AI integration in assessment. This relationship between frames and practice remains relevant even as the technology and teachers' understanding of it continue to evolve. Future research could benefit from longitudinal studies to track how these frames change over time and how such changes influence pedagogical practices.

Therefore, it is crucial to understand what teachers know, how they reason, and the decisions and assessment methods that result from this, their technological frames. Additionally, there is a need for teachers to receive ongoing professional development to create better conditions for an assessment practice that effectively incorporates AI as a tool. Future research and professional development initiatives could benefit from focusing on empowering teachers to engage more deeply in this form of conceptual and practice-oriented EUD, enhancing their ability to adapt and innovate in the face of rapidly evolving educational technologies. This could involve supporting teachers in moving towards the central intersection of the Venn diagram, where they can integrate AI as a learning objective, a tool, and a part of the assessment process in balanced and innovative ways.

References

[1] Bronwyn Cumbo and Neil Selwyn. 2022. Using participatory design approaches in educational research. International Journal of Research & Method in Education 45, 1: 60–72. https://doi.org/10.1080/1743727X.2021.1902981

[2] Pelle Ehn. 1989. Work-oriented design of computer artifacts. Lawrence Erlbaum, Hillsdale New Jersey.

[3] Mar Pérez-Sanagustín, Miguel Nussbaum, Isabel Hilliger, Carlos Alario-Hoyos, Rachelle S. Heller, Peter Twining, and Chin-Chung Tsai. 2017. Research on ICT in K-12 schools – A review of experimental and survey-based studies in computers & education 2011 to 2015. Computers & Education 104: A1–A15. https://doi.org/10.1016/j.compedu.2016.09.006

[4] Ben Williamson, Felicitas and John Potter. 2023). Re-examining AI, automation and datafication in education. Learning, media and technology, 48, 1: 1-5.

[5] Lin Phoebe and Jessica Van Brummelen. 2021, May. Engaging teachers to co-design integrated AI curriculum for K-12 classrooms. In Proceedings of the 2021 CHI conference on human factors in computing systems, 1-12.

[6] Gerhard Fischer. 2023. Adaptive and adaptable systems: Differentiating and integrating AI and EUD in: Spano D. (ed.) Proceedings of the 9th International Symposium on End-User Development, 3-18. Springer.

[7] Mike Sharples. 2022. Blog post: New AI tools capable of writing student essays compel educators to reconsider approaches to teaching and assessment. https://blogs.lse.ac.uk/impactofsocialsciences/2022/05/17/new-ai-tools-that-can-write-student-essays-require-educators-to-rethink-teaching-and-assessment/

[8] UNESCO. 2023. Guidance for generative AI in education and research. UNESCO

[9] Walter J Ong. 2002. Orality and literacy the technologizing of the word (2. ed.). London: Routledge.

[10] Wanda J Orlikowski and Debra C Gash. 1994. Technological frames: making sense of information technology in organizations. ACM Transactions on Information Systems (TOIS), 12, 2: 174-207.

[11] Wanda J Orlikowski. 2007. Sociomaterial practices: Exploring technology at work. Organization studies. 28, 9: 1435-1448.

[12] Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating Academic Answers Generated Using ChatGPT. Journal of Chemical Education, 100(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087