THE UNIVERSITY *of* EDINBURGH

AI-Driven Estimation and Enhancement of Metacognitive Monitoring to Improve Mathematical Learning in Children

Xingran Ruan

Doctor of Philosophy

EPCC

School of Informatics

The University of Edinburgh

2025

Abstract

Effective learners actively engage with and select their learning processes from metacognitive monitoring, a cornerstone of educational success. Metacognitive monitoring is an essential aspect of effective learning and enables learners to reflect on their thought processes, learning strategies, and knowledge states, thereby regulating their own learning. Enhancing metacognitive monitoring skills has shown substantial educational benefits, as evidenced by existing pedagogical and social research. This is particularly true in mathematics, where precise metacognitive monitoring correlates strongly with improved performance. Effective metacognitive monitoring is essential for enhancing mathematical abilities among learners.

Children aged 7 to 11 are in a pivotal developmental phase where metacognitive skills can be significantly shaped. This period is critical for cultivating these skills, which are vital for their future academic and personal growth. In traditional classroom settings, expert teachers encourage students to think about their own learning by asking them reflective questions. Yet, their time is limited, making it difficult to address each student's needs in large classes. With the progress of AI, computer-based learning environments (CBLEs) are increasingly capable of replicating learning support at scale and are well-positioned to tailor support in large, diverse learner populations. However, adapting teacher-driven interventions to CBLEs poses significant challenges. For example, intelligent tutoring systems (ITSs) often provide frequent prompts that can interrupt the natural flow of learning activities and may reduce learners' trust in these systems. Additionally, ITSs require learners to articulate their thoughts during self-reflection, a process that is essential yet complex. These subjective responses are often unreliable. To address these challenges, our work proposes a novel technique that estimates young learners' metacognitive monitoring performance (MMP) by analyzing their spontaneous facial responses, thus aiming to emulate the

nuanced approach of expert teachers within digital learning environments.

Building upon the prior work about emotion expressed during metacognitive monitoring, we developed the Meta-Facial Expression Interpreter (M-FEI), an approach to estimate MMP through facial cues. It enables real-time estimation and has been proven to outperform an existing conventional method. These conclusions have been derived from a first large user study conducted with 184 children aged 7 to 11 from two provinces in China and from Scotland.

An ITS designed to enhance learners' metacognitive monitoring was employed in a second large-scale user study. This compared mathematical learning outcomes between a tailored metacognitive monitoring intervention using M-FEI (condition 1) and, respectively, using the conventional method (condition 2). The study included a total of 215 children aged from 7 to 12. The results showed that children in condition 1 achieved improved learning outcomes and significantly outperformed those in condition 2.

This PhD thesis has pioneered an innovative approach for tailoring learning support by a multi-modality deep learning neural network. This approach has the potential to benefit a diverse range of learners for a variety of subjects by providing personalized and effective educational support in improving metacognitive monitoring skills.

Lay Summary

Thinking about your own learning is an essential skill for learning better. It occurs at various stages throughout the learning process, such as when learners assess their own knowledge of a task, evaluate their answers, or review the strategies they have used. To support this thinking process, various teaching methods have been developed, particularly in the field of mathematics. These methods help learners effectively check their answers and refine their own learning processes.

In real classrooms, expert teachers play a crucial role in noticing when children are thinking about their own learning. This is especially important for young learners who may struggle to explain what they think about their own work, such as being unsure whether their solutions are correct, or they may be over-influenced by their peers. Based on these insights, teachers guide learners to reflect on their own knowledge, helping them be aware of what they know and what they need to learn. Consequently, such support facilitates better learning outcomes among learners. Yet, their time is limited, making it difficult to address each student's needs in large classes.

In this research, we looked at how young students show their emotions and how they reflect on their own learning. We created a tool that looks at children's faces, where they are looking, and their head movements to identify how well a student is doing this. We called it the Meta-Facial Expression Interpreter, or M-FEI. It can give quick and reliable guesses about someone's self-reflection performance, and it did better than the older methods used in research. These conclusions have been derived from a big study with 184 children aged 7–11, carried out in schools in China and Scotland.

A smart computer program designed to help children reflect on their learning was used in a second big study. This study compared mathematical learning outcomes between a special support to help children reflect on their learning using M-FEI (con-

dition 1) and using the conventional method (condition 2). The study included a total of 215 pupils aged 7–12. The results showed that pupils in condition 1 achieved improved learning outcomes and did better than condition 2.

This PhD thesis introduced a new way to give children learning support. This approach has the potential to benefit a diverse range of learners for a variety of subjects by providing support that fits each child and really helps them learn.

Acknowledgments

I am very grateful for the trust and mentorship provided by my supervisors, Dr. Charaka Palansuriya and Dr. Aurora Constantin. I have learned a great deal from both of them and consider myself lucky to have been able to work with them. I have also enjoyed taking part in research groups at the Edinburgh Parallel Computing Centre (EPCC).

The EPCC and the wider University of Edinburgh are an impressive community. I benefited from the encouragement and support of Robin Hill and Mark Bull throughout my PhD research, and I am especially grateful to Nick Brown for his invaluable feedback, which helped shape and refine this thesis right up to the final moments before submission. In the background, Ben Morse, Jemma Auns, James Richards, and the rest of the administrative teams did a massive job of making my studies more enjoyable.

A number of people made my time in Edinburgh special. I would like to thank my colleagues Ricardo Jesus (my desk-mate), Mark Klaisoongnoen, Chao Tang, Weiyu Tu, Lilin Yu, Kejia Zhang, Jakub Adamski, Felicity (Flic) Anderson, Gabriel Rodríguez Canal, Ananya Gangopadhyay, David Kacs, Mateusz Meller, and Hovhannes Minasyan.

I am immensely grateful to Dr. Kangcheng Wang for his invaluable cooperation and expertise in conducting the user studies integral to this PhD thesis. His insights and dedication were pivotal in the success of my research.

I would also like to extend a heartfelt thank you to all the participants involved in the studies—the children and their parents or guardians. Their willingness to engage with my research made this work possible.

Special thanks are due to the principals and teachers from the schools that participated in my studies. Their support and assistance were crucial in facilitating the

smooth execution of my research activities associated with their institutions.

Additionally, I extend my deepest appreciation to Prof. Malcolm Atkinson, whose guidance has been invaluable since my collaboration began in April 2024. Prof. Atkinson has not only been a wise and kind mentor but also a supportive friend. His insightful ideas and constructive feedback have greatly enriched my PhD research, helping to shape it into a more comprehensive study. With his dedicated assistance, I successfully published a paper at CHI 2025 in Japan, a prestigious conference in the field of Human-Computer Interaction.

In the end, my parents, Weiqing Ruan and Binbin Tang, have provided constant support and encouragement. This PhD research is supported by my parents. I cannot imagine having done this without you.

Declaration

I declare that this thesis was composed by me, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Xingran Ruan

Contents

List of Figures

List of Tables

Critical Abbreviations

A

AD — Affective dynamics

C

CBLE — Computer-based learning environment

CHI — Computer-Human Interaction

F

FER — Facial emotion recognition

FOK — Feeling of knowing

H

HCI — Human-Computer Interaction

I

ITS — Intelligent tutor system

J

JOC — Judgment of confidence

JOL — Judgment of learning

K

KMA — Knowledge monitoring assessment

M

MASRL — Metacognitive and Affective Model of Self-regulated Learning

M-FEI* — M-FEI pairing with TimeSformer (K400) and AHG

ML — Machine learning

MMP — Metacognitive monitoring performance

MOOC — Massive Open Online Courses

S

SRL — Self-regulated Learning

Definitions

Metacognition: The broad definition of metacognition as 'thinking about thinking' is often interpreted widely (Winne 2011). Its definition includes any cognitive process that receives information from and has a controlling influence on another cognitive process (Shea et al. 2014).

Metacognitive monitoring: Metacognitive monitoring refers to evaluating the ongoing progress or current state of a particular cognitive activity (Winne 2011). In particular, the metacognitive monitoring in this PhD thesis refers to self-evaluating the ongoing task's solution.

Affective states: Affective states refer to the internal experiences associated with feelings, emotions, and moods.

Emotions: In this thesis, emotions are distinguished from 'affective states,' which are the formal, externally measurable aspects of affective states. They are externally expressed and recognizable responses, often identified through facial expressions.

Facial expressions: Facial expressions are visible movements of facial muscles that convey specific emotions.

Facial cues: They are more broadly, encompass these expressions as well as other nonverbal facial signals, including subtle movements such as gaze direction and head orientation. In this thesis, the facial cues are facial expressions, head gestures, and gaze directions.

Smart game/serious game: It is a computer-based game designed for a primary purpose other than pure entertainment.

Intelligent tutor system: An ITS is a computer-based platform designed to simulate human tutoring by enhancing learner engagement, improving retention, and supporting overall learning outcomes through immediate feedback and personalized instruction.

Chapter 1

Introduction

不积硅步，无以至千里；不积小流，无以成江海。(Translation: A journey of a thousand miles may not be achieved without the accumulation of each single step, just as the enormous ocean may not be formed without gathering every brook or stream.)

— 荀子 (Xun Zi)

This chapter presents an overview of the PhD thesis, including its background, motivation, prior research and its limitations, and the proposed approach to addressing these challenges. The idea behind this PhD thesis was presented at the ACM IDC 2022 Doctoral Consortium (DC) in a paper: 'Real-time Feedback based on Emotion Recognition for Improving Children's Metacognitive Monitoring Skill' (Ruan et al. 2022).

Learning is a dynamic process through which individuals acquire, process, and retain knowledge, skills, and attitudes. It involves the assimilation of new information and the integration of experiences into existing cognitive frameworks (Winne 2011). As learners engage with the learning environment, they exercise autonomy by selecting topics, choosing materials, and devising suitable learning strategies. Through self-regulation, they actively pursue and achieve predefined learning goals. Building on this foundation, self-regulated learning (SRL) further enhances learners' capabilities to manage and control their own educational experiences.

SRL encompasses key components such as goal setting, strategic planning, progress monitoring, and adaptive adjustments to learning strategies to optimize performance (Winne 2011). This learning skill is crucial for helping pupils achieve better learning

outcomes. However, noticeable differences exist in how effectively pupils develop and apply SRL skills. An example of this is the performance gap observed between students in mathematical learning outcomes, which highlights these disparities (Bullen et al. 2020, Siregar et al. 2020). Such performance gaps suggest that the SRL strategies in learners may be unevenly developed or inconsistently applied. To address this, expert teachers or tutors often play a crucial role by providing timely prompts and guidance to support learners in regulating their own learning processes.

Among the various components of SRL, metacognition, commonly defined as 'thinking about one's thinking', plays an essential role by enabling learners to evaluate their understanding, plan strategically, and adjust their approaches based on task demands (Schraw & Dennison 1994). This has been recognized as a crucial approach in improving learning outcomes (Higgins et al. 2016). Metacognitive processes encompass the strategic actions and the motivation required for effective learning (Winne 2011). Skilled self-regulated learners continuously monitor and regulate their cognitive activities, enabling them to evaluate the effectiveness of their learning strategies and make informed adjustments. Empirical studies have demonstrated that proficiency in metacognitive monitoring significantly enhances SRL and contributes to improved learning outcomes, particularly in complex subjects such as mathematics (Isaacson & Fujita 2006, Desoete & De Craene 2019, Sawyer et al. 2014, Brosnan et al. 2016, Grainger et al. 2016, Carpenter et al. 2019). For instance, a study by Isaacson & Fujita (2006), involving 84 students, found that learners who effectively monitored and evaluated their learning processes achieved higher final examination scores. Recently, this has emerged as an effective pedagogical approach to enhance learning outcomes through delivering tailored interventions to improve metacognitive monitoring performance (MMP) (Higgins et al. 2016).

Current interventions aimed at enhancing pupils' MMP generally fall into two categories. The first category, standard approaches, involves direct teacher guidance and structured problem-solving sessions, as documented in studies by Cogliano et al. (2020) and Montero et al. (2021). The second category includes computer-based interventions that use digital tools and interactive tutor systems to enhance metacognitive skills and simulate real-classroom teacher interactions, supported by findings from Maras et al. (2019) and Clabaugh et al. (2019). Detailed descriptions of these intervention types are provided in Section 2.4. Empirical studies demonstrate that both types of intervention effectively enhance learners' learning outcomes. For exam-

ple, Grawemeyer et al. (2015) reported improved grades among children who received metacognitive support, and Maras et al. (2019) noted significantly higher scores on mathematical questionnaires among supported groups compared to control groups. Similarly, studies by Cogliano et al. (2020) and Montero et al. (2021) observed improved performance following training with standard metacognitive interventions.

Compared to standard metacognitive interventions, computer-based interventions offer several additional benefits. Firstly, these digital platforms tend to create more engaging and comfortable learning environments, as children typically demonstrate increased enthusiasm and involvement when interacting with technology in secure settings (Valencia et al. 2019). Secondly, technological devices and advanced algorithms enable the capture and analysis of subtle, implicit indicators—such as facial expressions, head gestures, gaze directions, and body gestures—that can significantly enhance the quality of metacognitive support (Azevedo et al. 2022). Lastly, computer-based interventions facilitate automated personalization and consistent replication, allowing tailored feedback to be efficiently delivered to individual learners (Clabaugh et al. 2019).

Computer-based intervention designs often provide support through text-based feedback on past learning activities. For instance, in the approach described by Maras et al. (2019), immediate feedback is provided after each question attempt, reporting the current question's achieved score and the cumulative score. Subsequent feedback, delivered after completing a series of questions, includes the number of correct responses, reminders about the goal to maximize total scores, and strategic recommendations for managing increased difficulty. However, despite these notable advancements, there are significant limitations in existing computer-based interventions, especially for pupils.

Current mechanisms of computer-based interventions predominantly rely on learners' explicit self-rated responses to estimate their MMP to provide feedback (Kautzmann et al. 2016, Kautzmann & Jaques 2019). Therefore, learners are frequently required to assess their own performance, a repetitive and burdensome process. However, such mandatory self-assessment sessions have been shown to disrupt learners' flow and engagement in the learning process (Riku 2021). Moreover, subjective responses are often inaccurate, especially when pupils struggle to articulate their feelings (Brown et al. 2015), which undermines the effectiveness of subsequent interventions. Consequently, these inaccurate MMP estimations reduce the accuracy of computer-

ized interventions, sometimes providing unnecessary feedback to learners who do not need it while omitting feedback from those who require it. As a result, learners are at risk of experiencing frustration or disengagement. This issue becomes especially critical in scenarios where timely and responsive feedback is essential to maintaining learner engagement and sustaining learning momentum (Graesser 2020).

To address these limitations, an effective strategy involves leveraging objective data derived from learners' spontaneous behaviors to estimate their MMP. Prior research by D'Mello et al. (2009) investigated undergraduate students' behaviors during sessions interacting with computer-based metacognitive interventions. In these sessions, learners were encouraged to reflect more deeply on their learning strategy and received feedback for metacognitive monitoring. It was observed that the majority of the spontaneous behaviors were facial expressions, while other behaviors like head nods, shakes, and jaw movements were infrequent. Additionally, research has shown that occurrences of certain emotions, identifiable through facial expressions, are correlated with MMP in learning sessions (Taub & Azevedo 2018, Cloude et al. 2020, Taub et al. 2021). Notably, studies by Taub, conducted with undergraduate students, reveal that emotions such as boredom and surprise are significantly correlated with MMP during learning processes (Cloude et al. 2020, Taub et al. 2021). Even though the intricate interactions between learners' spontaneous behaviors and their MMP remain incompletely understood, prior research highlights potential directions for investigation.

Building upon the prior research, this PhD thesis aims to accurately estimate pupils' MMP by analyzing their spontaneous facial cues, including facial expressions, gaze directions, and head gestures. Using these estimations, this research tailors and provides metacognitive interventions that have validated benefits to MMP in previous studies, and tests the efficacy of these interventions on pupils' mathematical learning outcomes in a computer-based learning environment (CBLE).

To achieve this target, this PhD work addresses the following central research questions:

RQ1. To what extent does MMP impact pupils' learning outcomes in CBLEs, and can pupils' MMPs be inferred from their facial cues?

- RQ1.1 How does MMP influence pupils' task scores?
- RQ1.2 What facial cues have a significant correlation with MMP?

RQ1.3 What is the performance of the conventional approach in estimating pupils' MMP?

RQ1.4 Is it possible to estimate pupils' MMP using deep learning to interpret their facial expressions? If yes, does this improve MMP estimation?

RQ2. Given the established benefits of metacognitive interventions in educational research, can interventions tailored to MMP, as identified through facial cues, enhance pupils' mathematical learning outcomes in CBLEs?

RQ2.1 Does the intervention using the M-FEI approach improve pupils' mathematical learning outcomes?

RQ2.2 How do the mathematical learning outcomes of pupils who receive the tailored intervention using the M-FEI approach compare with those of pupils who undergo the KMA-based approach?

Addressing RQ1:

To investigate RQ1, this research initially experienced challenges due to the unavailability of relevant data. To address this, a smart game was developed, and a user study was conducted to collect the necessary data for analyzing pupils' MMP and their facial cues across different MMP states. The development of the smart game, the procedures of the user study, and the analysis of the collected data are detailed in Chapter 3.

Following this, the focus progressed toward estimating MMP using deep learning and facial expression interpretation techniques. Chapter 4 presents a comparison between conventional approaches and the newly proposed Meta-Facial Expression Interpreter (M-FEI). Experimental results illustrate the advantages of M-FEI in terms of performance, generalizability, and educational applicability, which are analyzed in detail in the same chapter.

Addressing RQ2:

To address RQ2, an additional user study was conducted to validate the educational benefits of the M-FEI approach, as detailed in Chapter 5. Although this research focused on the context of mathematics, metacognitive monitoring is a universal learning skill applicable across various educational activities. Therefore, it is anticipated that the proposed contributions could be extended to other learning contexts in future work, which is highlighted in Chapter 6.

1.1 Thesis Structure

The following provides a structured overview of this thesis:

Chapter 2 presents the foundational background and related research underpinning this study. It begins with an examination of the expanding role of artificial intelligence and its potential within education. Subsequent sections explore validated interventions for enhancing metacognitive monitoring, which are grouped into standard and computer-based approaches, highlighting educational methodology advances driven by technology. The chapter also explores the concept of metacognitive monitoring within SRL, emphasizing its significance for educational success and detailing how it can be effectively supported in children. Finally, this chapter reviews available techniques for estimating MMP, specifically focusing on interpreting facial expressions within educational settings.

Chapter 3 details a user study that examines the collection of facial cues associated with pupils' metacognitive monitoring. This research involves developing an application, called Meta-Brainhood, designed to stimulate metacognitive monitoring while capturing relevant facial cues. The results observed in the user study provide additional support for the impact of MMP on learning performance in a CBLE. Findings from this study identify facial cues significantly correlated with the MMP of pupils, indicating their potential as real-time indicators in educational contexts.

Chapter 4 outlines the development of a deep-learning model, Meta-Facial Expression Interpreter (M-FEI), aimed at accurately estimating MMP through facial expression analysis. Leveraging data from a large-scale user study detailed in Chapter 3, Chapter 4 addresses the limitations of conventional methods, which often inaccurately evaluate MMP. The objective is to refine estimation accuracy, enabling more effective and minimally intrusive educational interventions for pupils.

Building on the technique developed in Chapter 4, Chapter 5 investigates the implementation and educational effectiveness of the intelligent tutor system (ITS) designed to enhance pupils' metacognitive monitoring. This system employs both the novel M-FEI-based approach and a conventional method, providing tailored interventions aligned with learners' estimated MMP.

Chapter 6 concludes the key findings of the research questions and the key contributions of this research. It also reports limitations and future works in this research.

1.2 Main Contributions

Throughout this PhD project, a methodology was successfully developed and implemented to improve young pupils' mathematical learning outcomes through metacognitive interventions tailored based on facial interpretation. Specifically, the research investigated the feasibility of estimating MMP through facial cues and validated the effectiveness of M-FEI-based metacognitive monitoring interventions in enhancing pupils' mathematical performance in a CBLE.

This work not only confirms the practicality of using facial expressions to tailor metacognitive monitoring interventions but also reveals the potential of real-time interventions to significantly improve educational outcomes. Furthermore, the methodology established in this research provides a framework for other researchers aiming to integrate deep-learning-based techniques into educational technologies. This methodology encompasses the design of data collection studies, data pre-processing, model proposals, and the validation of educational benefits associated with these models.

1.3 Publications Resulting from This Research

It should be noted that some content of this PhD thesis has been published in peer-reviewed conferences during this PhD research. Their relation to this thesis will be mentioned at the beginning of the relevant chapters. The following publications are available online.

1. Ruan, X., Palansuriya, C. and Constantin, A. (2022), Real-time Feedback based on Emotion Recognition for Improving Children's Metacognitive Monitoring Skill. In: *Proceedings of the 21st Annual ACM Interaction Design and Children Conference*, pp. 672–675. DOI: https://doi.org/10.1145/3501712.3538831.

2. Ruan, X., Palansuriya, C., Constantin, A. and Tsiakas, K. (2023), Supporting Children's Metacognition with a Facial Emotion Recognition-based Intelligent Tutor System. In: *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, pp. 502–506. DOI: https://doi.org/10.1145/3585088.3593882.

3. Ruan, X., Palansuriya, C. and Constantin, A. (2023), Affective Dynamic Based Technique for Facial Emotion Recognition (FER) to Support Intelligent Tutors

in Education. In: *International Conference on Artificial Intelligence in Education*, Springer, pp. 774–779. DOI: 10.1007/978-3-031-36272-9_70.

4. Ruan, X., Wang, K., Palansuriya, C. and Constantin, A. (2024), Identifying Children Metacognitive Monitoring Performance Through Facial Expressions. In: *Companion Proceedings of the 2024 Annual Symposium on Computer-Human Interaction in Play*, pp. 242–248. DOI: https://doi.org/10.1145/3665463.3678790.

5. Ruan, X., Constantin, A., Palansuriya, C., Wang, K., and Atkinson, M. (2025), Nurturing Self-aware Learning through Facial Expression Interpretation. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1-8 DOI: https://doi.org/10.1145/3706599.3720258.

Chapter 2

Background of Research

学如逆水行舟，不进则退。(Translation: Learning is like rowing upstream; not to advance is to drop back.)

— **曾国藩** (Zeng Guofan)

2.1 Artificial Intelligence in Education

Artificial Intelligence (AI) is often stereotypically associated with advanced robotic systems endowed with vast computational power and adaptive behaviors that emulate human cognition. However, in the educational domain, AI applications have evolved far beyond these traditional interpretations (Chen et al. 2020). Modern AI systems are increasingly employed to assess student work, deliver real-time feedback, and pinpoint areas for improvement (D'Mello & Graesser 2013). Moreover, these technologies facilitate the creation of virtual learning environments that foster interactive and immersive educational experiences (Guo 2020).

The integration of AI technologies in the educational setting has increased markedly in recent years, garnering significant scholarly and practical attention. For instance, Google's DeepMind has investigated methods to enhance personalized learning through the implementation of AI-based learning systems (Tombazzi et al. 2023). Similarly, the German Research Center for AI (DFKI) has established a dedicated research group that focuses on the intersection of educational technology and AI applications in learning and teaching (*German Research Center for Artificial Intelligence (DFKI)* 2025).

These developments highlight a growing recognition of AI's potential to transform educational practices through more adaptive and personalized learning experiences.

Integrating AI into educational platforms enables the correlation of learners' behavioral data with their ongoing learning performance. AI-driven analysis facilitates the evaluation of individual learning competencies by inferring students' thought processes and capabilities. For example, Orji et al. (Orji & Vassileva 2022) employed AI techniques to classify learner characteristics, thereby allowing for dynamic, personalized interventions that adapt to each student's needs. As educational paradigms evolve, AI systems are increasingly prepared to offer support grounded in established tutoring models and their inherent pedagogical frameworks. Additionally, modern user interfaces that accommodate various input modalities, such as voice, text, and clicks, provide multifaceted feedback through diverse output formats, including text, graphical representations, and visual cues. Advanced human-computer interaction functionalities, such as natural language processing, speech recognition, and affective state detection, further augment the capacity of these systems to enhance learning outcomes (Graesser 2020).

2.2 Mathematical Learning Interventions

Interventions aimed at enhancing learners' mathematical performance are typically classified into two categories: standard and computer-based. Standard interventions are most frequently employed in educational research and delivered by teachers or tutors in classrooms. These approaches use visual aids, such as images, diagrams, and numerical equations, to facilitate learners' self-reflection in problem-solving (Fang 2012, Jitendra & Star 2011). Empirical evidence suggests that participants benefit from these interventions and demonstrate significant improvements in mathematical learning outcomes, notably achieving a deeper understanding of the underlying mathematical structures (Herzog & Casale 2022).

In contrast, computer-based interventions employ various technological approaches, including tablet-based video modeling, interactive software interfaces, and robotic assistants (Jowett et al. 2012, Grawemeyer et al. 2015, Clabaugh et al. 2019). These approaches offer the advantage of scalable and consistent delivery, overcoming limitations associated with the traditional learning environment, such as limited teacher availability, challenges in providing individualized support, and the difficulty of adapt-

ing interventions across large-scale educational settings. For example, Clabaugh et al. (2019) deployed a Socially Assistive Robot (SAR; see Figure 2.1) in 17 households to support primary arithmetic practice. The SAR was programmed to dynamically adjust both task difficulty and the intensity of feedback in response to the number of errors committed by the learner. Specifically, if a child exceeded a predetermined error threshold, the difficulty level was reduced; otherwise, it was maintained or increased. In addition, corrective feedback was provided immediately upon an error being made by the child, with its level calibrated to minimize subsequent mistakes while remaining within acceptable limits. Empirical results demonstrated that participants exhibited significantly higher scores in numerical operations and mathematical reasoning following this intervention compared to baseline measures.

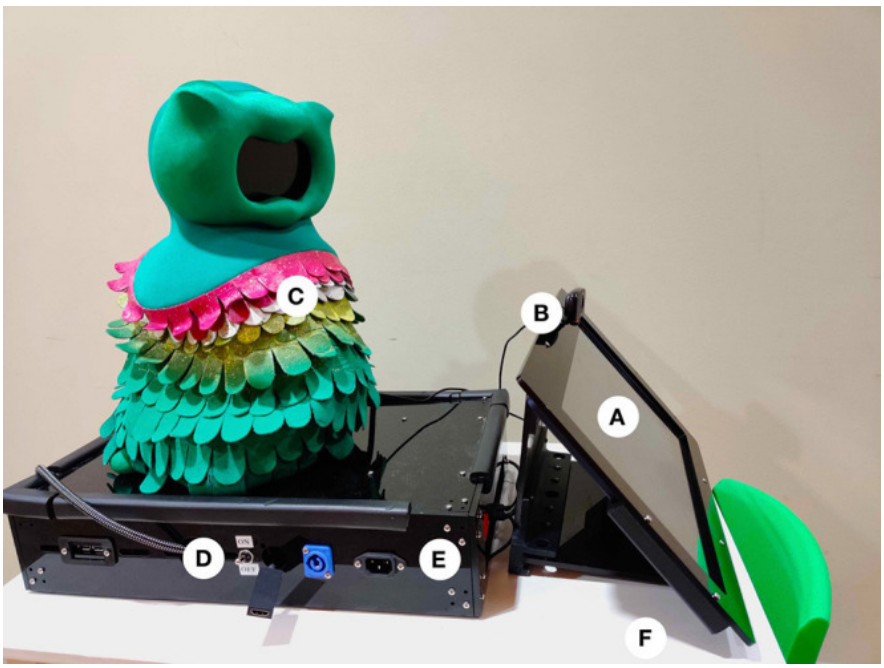


Figure 2.1: Example physical in-home learning environment (Clabaugh et al. 2019).

Grawemeyer et al. (Grawemeyer et al. 2015) evaluated the effectiveness of an open learner model (OLM), the Maths Island Tutor (see Figure 2.2), in enhancing the understanding of mathematical concepts among 24 children. The experiment was structured into three sessions: in session 1, no errors were present in the conceptual representations; in session 2, errors were introduced without informing the participants; and in session 3, errors were introduced and explicitly communicated. Following these sessions, participants were instructed to identify the mathematical concepts they had learned by selecting the corresponding flags on the mathematics

islands and by circling the flag associated with any error (see Figure 2.2). The results indicate that the use of an OLM can foster metacognitive benefits in young learners.

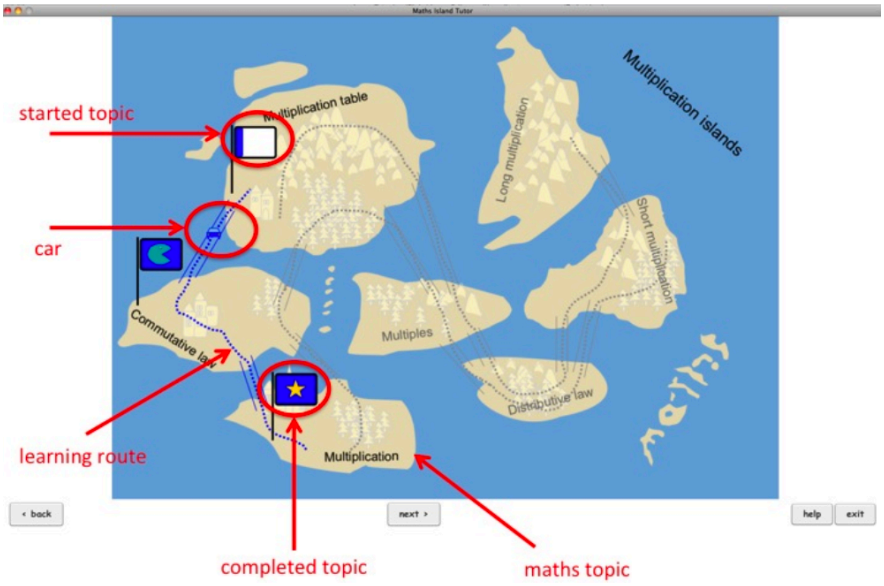


Figure 2.2: Example interface of Open Learner Model (OLM) (Grawemeyer et al. 2015).

Maras et al. (Maras et al. 2019) introduced a novel computer-based metacognitive intervention, the Math Challenge, to assess the effectiveness of feedback in enhancing mathematical learning. The Math Challenge is structured into seven progressively challenging levels, with increasing question complexity at higher levels. The program commences with comprehensive instructions outlining its structure, objectives, and point allocation system. Prior to each problem, a five-point pre-test intention measure is administered to gauge the learner's anticipated effort in answering correctly. Following each question, a five-point post-test confidence judgment was obtained, after which immediate feedback was provided, detailing the correctness of the response, the points earned for that item, and the cumulative score. Upon completion of each level, additional summary feedback was then delivered, comprising the number of correct responses, total points accumulated, a reminder of the learning goal, and strategic recommendations (e.g., lowering the difficulty level results in simpler questions but fewer points per item). Empirical findings indicated that participants who received feedback achieved significantly higher scores on the Math Challenge compared to those who did not.

Within computer-based interventions, the way support is delivered by intelligent agents has become increasingly sophisticated. Intelligent Tutoring Systems (ITSs) are designed to respond not only to learners' answers, are designed to respond not only to learners' answers, but also to a variety of behaviors, such as gestures, facial expressions, spoken language, and even physiological signals like heart activity (ECG) and brain activity (EEG) (Graesser 2020).

One well-known example is AutoTutor D'Mello & Graesser (2013), which was created to engage learners in conversational interaction. Instead of simply giving correct answers, AutoTutor guides learners through supportive dialogue, for instance, by prompting with questions like 'Could you at least try to give me an answer?' or offering encouragement such as 'Let's try this together.' Through these conversational moves, AutoTutor motivates learners to elaborate on their ideas and express them in complete sentences or explanations.

In addition, other ITSs have sought to interpret learners' nonverbal signals, such as facial expressions and body gestures, in order to provide more adaptive and context-sensitive support. An affective-sensitive AutoTutor (Azevedo et al. 2009) was implemented with 256 production rules that specified how AutoTutor was to respond to the learner's emotions. For example, if a learner is not performing as expected, AutoTutor adapts by adjusting the content to sustain engagement. To challenge high-performing learners, it introduces more difficult problems. When a low-performing learner is frustrated, AutoTutor offers a hint or an easier question to guide the learner back on track.

2.3 Metacognitive Monitoring

The process of evaluating the process of learning or the current state of knowledge is referred to as metacognitive monitoring (Rivers et al. 2020).

2.3.1 Role of Metacognitive Monitoring in Learning

Self-regulated learning (SRL) is the intentional and strategic adaptation of learning activities to achieve specific learning goals (Winne & Hadwin 2010), and is commonly practiced by successful learners. Winne (2011) elucidates the interdependent roles of metacognition in the SRL. Throughout the phases of SRL, learners have opportuni-

ties to metacognitively monitor properties of information, declarative and procedural knowledge, and their cognitive experience.

To emphasize the role of metacognitive monitoring in SRL, this subsection outlines the dynamic sequence of four distinct phases and illustrates how metacognitive monitoring operates within them. In Phase 1 (Defining the Task), learners begin by assessing their understanding of the task at hand and may engage in metacognitive monitoring to clarify task requirements. As they start working, they might also adopt metacognitive control strategies if they recognize that the task is unclear. Moving into Phase 2 (Setting Goals and Planning), learners implicitly evaluate possible goals and plan cognitive strategies to achieve them. During Phase 3 (Engagement), learners become actively involved in the task, applying tactics, constructing outputs, and engaging in cognitive processes, while metacognitive activities occur spontaneously and alongside their cognitive efforts. Finally, in Phase 4 (Large-Scale Adaptation), learners reflect on their overall approach after completing the task, aiming to develop strategies that will make handling similar tasks more effective in the future. Throughout these phases, metacognitive monitoring plays a crucial role by allowing learners to continuously assess their understanding, track their progress, and make informed adjustments to their strategies, thereby supporting effective self-regulated learning.

This process can be illustrated with a practical example. When learners are asked the question, 'What caused the fall of Rome?', they engage in metacognitive monitoring by continuously evaluating the completeness and adequacy of their responses. Specifically, they assess the quality of the information recalled, the effectiveness of the strategies used to retrieve and organize that information, and their overall confidence in the response. Similar monitoring occurs in mathematics, such as when a child solves '$245 + 378 = ?$' and then reflects on whether their calculation seems correct, or when working through a multi-step word problem and checking if each step follows logically. In reading comprehension, learners may pause after a passage to consider whether they have understood the main idea, or in science, they may review whether their explanation of why plants need sunlight is accurate. Across these tasks, learners monitor the quality of their reasoning and their confidence in their responses.

2.3.2 Measurements for Metacognitive Monitoring Performance

Previous studies have investigated the measurements of learners' ability to monitor their cognitive processes. For instance, Wojcik et al. (Wojcik et al. 2013) presented children with 20 pairs of nouns—each pair consisting of a 'cue' and an 'answer.' Upon receiving a cue, children provided a Feeling of Knowing (FOK) judgment by indicating 'yes' if they anticipated recognizing the associated answer or 'no' if they did not. Subsequently, participants selected an answer corresponding to the cue, and MMP was quantified by the degree of agreement between their FOK judgments and the actual responses.

In a related study, Grainger et al. (Grainger et al. 2016) evaluated MMP in 64 children by having them watch a four-minute educational video, after which they completed a worksheet on its content. Once all questions were answered, the children provided a Judgment of Confidence (JOC) for each response on a seven-point scale ranging from 'extremely unsure' to 'extremely sure.' The researchers calculated the average confidence for correct responses and incorrect responses, defining the difference between these averages as an index of metacognitive monitoring accuracy. Similarly, Brosnan et al. (Brosnan et al. 2016) employed a comparable experimental design by administering a set of mathematical questions followed by a JOC worksheet to assess children's MMP. These results from comparing MMP indicate that children's ability to accurately monitor cognitive processes is notably limited, especially when tasks are complex or time-constrained (Brosnan et al. 2016, Grainger et al. 2016). Researchers found that MMP among children appears to be heterogeneous, varying with task complexity and the conditions under which tasks are performed. Specifically, when confronted with more demanding or strictly timed tasks, children tend to exhibit less precise metacognitive monitoring.

2.4 Metacognitive Interventions for Math Learning

Supporting metacognition is recognized as one of the most effective and cost-efficient educational interventions (Desoete & De Craene 2019). Research in educational psychology has consistently demonstrated that metacognitive abilities are strong predictors of academic achievement, often surpassing the predictive power of intellectual capabilities (Isaacson & Fujita 2006, Desoete & De Craene 2019). Furthermore,

extensive evidence underscores the effectiveness of metacognitive development as a strategic intervention for enhancing the educational outcomes of school children (Higgins et al. 2016, Montero et al. 2021).

This section examines a range of educational interventions designed to improve students' mathematics performance by enhancing their MMP. In line with interventions used in mathematics education (see Section 2.2), metacognitive interventions are introduced in two categories: standard and computer-based.

2.4.1 Standard Metacognitive Monitoring Intervention

Standard metacognitive interventions are delivered by teachers in traditional classroom settings, often involving guided questioning, feedback, and structured reflection activities.

Montero et al. (2021) conducted a study to investigate whether metacognitive training could enhance the mathematics performance of students aged 12 to 13 (Grade 7). The study involved 40 students (19 males and 21 females) from a high school in the Philippines. Initially, the students' metacognitive monitoring skills were assessed using the Metacognitive Awareness Inventory (MAI) (Sperling et al. 2012), a tool with an 18-item self-report questionnaire designed to evaluate metacognition in children from 3rd to 9th grade. The participants then underwent six weeks of metacognitive training from October 3 to December 2, 2016, including strategy instruction, guided metacognitive interventions, structured practice, and performance evaluation. This training was integrated into their Grade 7 mathematics lessons using a method known as I.M.P.R.O.V.E (Mevarech & Kramarski 1997), covering the entire unit on Algebra. Post-intervention, a second MAI assessment was administered to evaluate any changes in the student's metacognitive skills, particularly in mathematics. The effectiveness of the intervention was analyzed using a t-test[1] to compare the pre-test and post-test MAI scores and use `Cohen's d` (Cohen 2013) to measure the significance of the changes. The results indicated a significant improvement, with a t-value of 5.983 and $p<.05$[2], suggesting that the metacognitive training had a substantial impact. Among the metacognitive processes, monitoring, planning, and evaluating, monitoring

[1] The t-test is a statistical method used to determine whether there is a significant difference between the means of two groups.

[2] A t-value of 5.983 indicates a large difference between group means relative to the variability in the data, and $p<.05$ indicates that this difference is statistically significant at the 5% level.

showed the most significant improvement, with a `Cohen's d` value of 1.00[3], indicating a significant difference. Despite these positive outcomes, the study had limitations, notably the absence of a control group. This omission makes it difficult to conclusively attribute the improvements solely to the metacognitive training, as other factors such as natural maturation or learning from other courses could also influence the results. Furthermore, while the participants, both students and teachers, provided positive feedback on the intervention, the lack of a control group remains a significant drawback in the research design.

Cogliano et al. (2020) investigated the impact of training in metacognitive monitoring and control skills on undergraduate students' academic performance and the accuracy of their metacognitive judgments. The study involved 103 undergraduate students who initially completed a pre-test using the Metacognitive Awareness Inventory (MAI) method. Following this, the students participated in ten weekly multiple-choice practice tests, which included feedback. They were also required to complete feedback assignments that prompted them to self-assess their understanding of well-learned versus yet-to-be-learned material based on the feedback received. Furthermore, the students evaluated the effectiveness of their study strategies and planned their future study approaches. By the third week, the students were randomly assigned to either a trained group or a control group. The trained group received instruction on the benefits of retrieval practice, learned how to self-regulate their learning strategies, and evaluated the feedback independently. Meanwhile, the control group engaged in reading and activities relevant to the course content but not focused on metacognitive skills. The outcomes of the study were assessed through course examinations conducted in Weeks 6, 10, and 15. The results demonstrated that students in the trained group performed better on final exam questions that had not been covered in the quizzes compared to those in the control group. This suggests that metacognitive training, particularly in retrieval practice and feedback evaluation, can significantly enhance students' ability to manage their learning and improve their academic performance.

[3] `Cohen's d` is a measure of effect size that quantifies the magnitude of difference between two group means, expressed in standard deviation units. A larger value indicates a stronger effect, with common benchmarks being 0.2 for a small effect, 0.5 for a medium effect, and 0.8 or above for a large effect (Cohen 2013).

2.4.2 Computer-based Metacognitive Monitoring Intervention

Computer-based metacognitive interventions are implemented within computer-based learning environments (CBLEs), using digital tools or intelligent systems that align with theories to assist learners in monitoring and regulating their thinking.

Math Challenge, which is proposed in Maras et al. (2019), stands out as a computer-based intervention designed to enhance children's mathematics performance by improving their metacognitive monitoring accuracy. This intervention features a structured program with seven ascending levels of difficulty, with each level presenting increasingly challenging questions. To aid in developing the children's metacognitive monitoring skills, the Math Challenge provides immediate feedback after each question. This feedback includes the total score accumulated, the correct answer to the current question, the overarching learning objective, and a preview of the challenges expected at the next level. Such detailed feedback is crucial in helping children understand their learning process and adjust their strategies accordingly. While the study by Maras et al. did not specifically measure the changes in metacognitive monitoring accuracy before and after the intervention, it did evaluate the overall effectiveness of the program by comparing the average mathematics scores between the feedback group and a control group. The results indicated a significant difference in mathematical performance, with the feedback group achieving an average score of 7.67, markedly higher than the control group's average of 3.21. This outcome suggests that the Math Challenge effectively supports children in enhancing their mathematical abilities, likely facilitated by the metacognitive insights gained through the structured feedback provided.

Kautzmann et al. (Kautzmann & Jaques 2019) developed an Animated Pedagogical Agent (APA) designed to foster metacognitive engagement among learners within a CBLE. The APA provides learners with four levels of metacognitive prompts based on estimations of their MMP. At the first level, the prompts are more abstract, encouraging students to activate their prior knowledge by reflecting on the task statement—for example, 'Before entering a new step for the equation you are solving, try to identify the parts of the equation and think about your knowledge to solve it.' The second level prompts help students consider the knowledge they have already applied in previous tasks, guiding them with statements such as 'Before you proceed, I want you to reflect on your knowledge. I think you have the knowledge to take a correct

new step in the equation. And you, what do you think about that?' At the third level, the prompts direct students to recall and connect with tasks similar to the current one they have previously solved, using prompts like 'Try to identify the parts of the equation and think if you have already applied any knowledge you can use now.' Finally, at the fourth level, the APA utilizes the ITS interface to display similar steps the student has successfully completed before, prompting them to draw from their prior solutions: 'I have selected steps that you had previously solved. Think about your past solutions to identify if you have the knowledge to solve the current step.' The results of this intervention demonstrated that pupils who completed exercises with these metacognitive prompts achieved significantly higher mathematical performance compared to those who did not receive such support.

By comparing standard metacognitive interventions with computer-based interventions, it becomes clear that traditional approaches rely heavily on teachers to deliver interventions within classroom settings. Teachers guide and encourage learners to reflect on their learning; however, their time and attention are limited, making it challenging to address individual student needs, particularly in large classes. Computer-based interventions help mitigate this limitation by providing scalable learning support through educational software, intelligent tutoring systems, and adaptive learning platforms. These systems can deliver consistent, tailored metacognitive support, making them a promising solution for enhancing self-regulated learning in diverse educational environments.

2.5 Estimation of Metacognitive Monitoring Performance

This section introduces the estimation of MMP, a central component of this research that will be further developed and implemented in detail in Chapter 4 and Chapter 5. Accurate estimation of MMP is essential for tailoring metacognitive support to learners in CBLEs, enabling timely and personalized interventions that promote more effective learning processes.

2.5.1 Conventional Approach of Estimating Metacognitive Monitoring Performance

In the conventional approach, MMP is estimated based on the agreement between learners' JOC and their actual performance on previous tasks (Kautzmann et al. 2016, Kautzmann & Jaques 2019). In this method, learners are asked to explicitly rate how confident they are in the correctness of their solutions after completing each task, and these self-reported confidence ratings are then compared to their actual outcomes. For example, the APA, which was developed in (Kautzmann & Jaques 2019), prompts pupils to articulate their confidence right after a task they solved. The system employs a Knowledge Monitoring Assessment (KMA) matrix (Romesburg 2004), which repeatedly measures an estimation of a learner's MMP. The KMA leverages the Hamann coefficient (Romesburg 2004) to statistically correlate task scores with self-rated confidence levels[4]. This measure provides insight into the alignment, or misalignment, between perceived and actual performance. By continuously monitoring this alignment, the APA is capable of adapting the subsequent metacognitive support in real-time, tailoring its feedback to address each learner's specific needs based on previous monitoring performance.

Similarly, Guo et al. (Guo 2020) adopted a variant of the KMA within a serious gaming environment to estimate MMP. In their approach, the system relies on historical MMP data to determine the appropriate information prompts while learners engage with game-based tasks. This method of using past performance data allows the system to deliver context-aware, adaptive metacognitive support, ultimately aiming to enhance learning outcomes.

However, this conventional approach relies on repetitive self-assessment questionnaires, requiring learners to frequently evaluate their performance. Such activities have been shown to disrupt learners' flow and engagement (Riku 2021), and the subjective nature of these responses often results in inaccurate evaluations.

[4]This is an important measurement of MMP. This PhD thesis will introduce this method in Section 4.6.1.

2.5.2 Potential Approach of Estimating Metacognitive Monitoring Performance

In addition to cognitive assessments, emotions play a significant role in learning (Efklides 2011). Recognizing this, researchers have increasingly explored the relationship between changes in emotions during learning activities and metacognitive monitoring accuracy (Taub et al. 2018, Taub & Azevedo 2018). This line of inquiry is predicated on the understanding that emotions can alter cognitive processes such as attention, memory retrieval, and problem-solving abilities, all of which are integral to learning. Changes in emotion during learning can either impact or reflect metacognitive processes. For example, positive emotions like interest and happiness might boost cognitive resources and promote deeper engagement with the learning material, thereby enhancing metacognitive accuracy (Taub et al. 2021). Negative emotions such as frustration or boredom might reflect cognitive overload or distraction, showing one's inability to accurately self-assess their learning (Taub et al. 2021).

2.5.2.1 Relationship Between Experienced 'Feelings' and Metacognitive Monitoring

As discussed in Jack et al. (2012), the concepts of emotion, facial expressions, facial cues, and affective states are often conflated between research fields. As these distinctions are critical to the framework of this thesis, this thesis clarifies them here. Affective states refer to internal experiences associated with feelings, emotions, and moods. Emotions are the formal, externally measurable aspects of affective states, expressed through observable responses and shaped by subjective interpretation. Facial expressions are visible movements of facial muscles that convey specific emotions. Facial cues, more broadly, encompass these expressions as well as other nonverbal facial signals, including subtle movements such as gaze direction and head orientation.

The complex relationship between learners' affective states[5] and their metacognitive monitoring process during the learning process has been explored in educational research. Several studies emphasize the bidirectional interplay between metacognitive activities, such as evaluating, monitoring, and planning, and learners' affective states, indicating that each can significantly shape the other. According to Mandler

[5]Here, we use the term affective states, as the remainder of the research is grounded in theoretical analysis and focuses on describing learners' feelings.

(Mandler 1989), the metacognitive monitoring process can provoke visceral arousal and spontaneous affective responses. This reaction typically occurs when learners encounter challenges or inconsistencies that disrupt their cognitive equilibrium. These disruptions can lead to feelings of frustration or confusion, which are common when learners realize their understanding is not as profound as they assumed (Mandler 1989, D' Mello & Graesser 2012). Moreover, Efklides (Efklides 2011) proposed the MASRL model, which describes the dynamic interactions between metacognition and affective states during SRL. The MASRL model suggests that both metacognitive experience and affective states are influenced by task features and task processing. For instance, the fluency of processing or occurrences of cognitive interruptions can simultaneously influence both affective states and metacognitive experiences, indicating that the experienced affective state reveals the metacognitive process (Efklides 2011).

2.5.2.2 Relationship Between Emotions and Metacognitive Monitoring

In a study by Cloude et al. (Cloude et al. 2020), the interplay between expressed emotions and metacognitive monitoring accuracy was examined in a study of 117 undergraduate students. The study used iMotions (iMotions 2018) technology[6], which measures emotions such as confusion, boredom, and frustration at several intervals —specifically at 0, 14, 28, and 42 minutes into the learning session. MMP for each participant was quantitatively assessed using a weighted formula: 50% of the correlation between the JOL scores and actual quiz performance, 25% from the quiz scores themselves, and the remaining 25% from the JOL scores. This approach provided a nuanced view of each student's ability to accurately gauge their learning. The results of the study highlighted a significant negative relationship between increases in boredom and MMP, indicating that higher levels of boredom were associated with decreases in metacognitive accuracy. Additionally, persistent confusion was found to have a detrimental effect on performance, suggesting that ongoing uncertainty impairs effective learning and metacognitive evaluation. This implies that learners exhibiting higher levels of boredom are likely experiencing lower engagement and, consequently, unsatisfactory metacognitive monitoring. These preliminary findings have profound implications for revealing MMP by interpreting facial expressions in metacognitive

[6]iMotions is employed to measure emotion by generating a vector of evidence scores corresponding to various emotions.

monitoring processes.

Taub and colleagues (Taub et al. 2021, 2018) have contributed significant empirical research that explores the relationship between specific emotions and MMP in undergraduate students. Their studies reveal that different emotions distinctly influence MMP, with notable implications for educational strategies and interventions. Their research found that surprise is negatively correlated with MMP. This suggests that when learners encounter unexpected information or outcomes within the learning process, it often leads to a disruption in metacognitive monitoring. This disruption could stem from the cognitive recalibration required when new information contradicts prior knowledge or expectations, leading to decreased satisfaction with one's metacognitive assessments. Conversely, the emotion of frustration, although typically viewed negatively, showed a positive correlation with MMP according to Taub's findings. This counterintuitive result suggests that frustration might act as a motivational force that pushes learners to address and overcome challenges, thereby enhancing their metacognitive monitoring. Frustration might stimulate deeper engagement with the task and more thorough exploration of different problem-solving strategies, ultimately leading to improved metacognitive evaluations.

These insights highlight the complex interplay between emotions and metacognitive processes, offering a new perspective for analyzing MMP in educational settings. The revealed correlation in undergraduate students suggests a potential approach to interpret pupils' MMP through spontaneous emotional and behavioral responses that naturally occur during learning. Compared to MMP estimation based on JOC, which relies on repetitive and burdensome self-reports, this potential enables more immediate and objective insights into learners' metacognitive states.

2.6 Computational Approaches for Emotion Interpretation in Education

2.6.1 Foundations of Machine Learning and Deep Learning

2.6.1.1 Support Vector Machine Emotion Classifier

Machine learning (ML) and deep learning provide computational methods for identifying patterns in data and making predictions. In emotion interpretation, Support

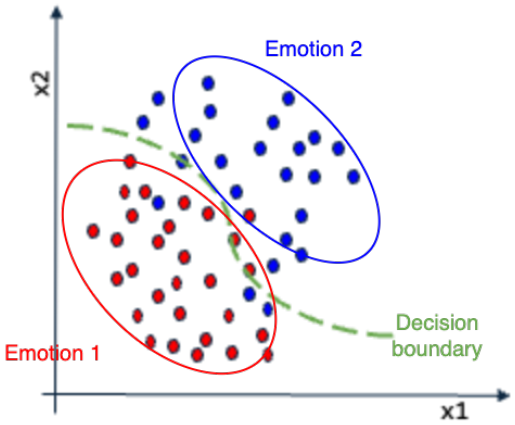


Figure 2.3: Example support vector machine (SVM) classifier for identifying emotions

Vector Machines (SVM) are widely used classical ML algorithms that classify data by finding the optimal hyperplane separating different classes (Hearst et al. 1998). This approach is particularly effective for structured, low-dimensional features such as facial expressions, gaze directions, and head gestures.

As illustrated in Figure 2.3, SVM aims to find the boundary (green line) maximizing the distance between classes (i.e., red and blue), improving generalization to unseen data (Hearst et al. 1998). With the boundary established from the training data, the SVM classifier classifies new data samples (for example, pupils' facial cues) by determining on which side of the boundary (which class of emotion) they fall.

2.6.1.2 Deep learning Neural Networks Emotion Classifier

Deep learning (LeCun et al. 2015), as a subset of ML, leverages multi-layered neural network structures to automatically learn hierarchical representations from raw data. In this process, the initial layers typically extract low-level features such as edges, colors, and simple textures from the input images. Subsequent layers build upon these to detect more complex patterns like facial components (e.g., eyes, mouth, eyebrows) and their configurations. Deeper layers combine these patterns to form high-level abstract representations associated with emotional expressions. Through this hierarchical processing, deep learning models can progressively transform raw visual data into meaningful features that support emotion interpretation without requiring manual feature design.

Figure 2.4 shows an example of a convolutional neural network (CNN) (Cao et al.

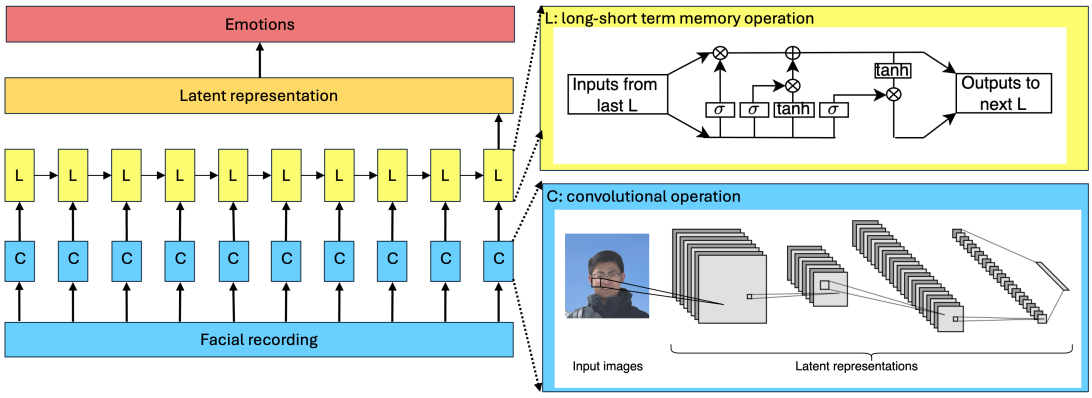


Figure 2.4: Example convolutional neural network for identifying emotions. Note: σ and $tanh$ are two functions for non-linear operations on input values (Dubey et al. 2022).

2018) for identifying emotion from an input facial recording. The deep learning model can automatically extract subtle visual cues, such as a slight frown or raised eyebrows. The following layers combine these representations across frames, and a Long Short-Term Memory (LSTM) network (Singh et al. 2023) is then used to capture the temporal dependencies between these cues over time. This allows the model to understand how emotional expressions evolve throughout a sequence, rather than treating each frame independently. By learning these patterns directly from large datasets of pupils' facial expressions, the model becomes capable of interpreting emotions across diverse individuals.

2.6.2 Leveraging Technology for Interpreting Emotions in Educational Contexts

Understanding learners' emotional states has been studied in prior work. With advancements in artificial intelligence, particularly in deep learning, it has become increasingly feasible to detect emotions by analyzing facial expressions. This section introduces existing research and technological developments that leverage AI-based approaches to interpret emotions through facial expression analysis, laying the groundwork for their application in educational contexts.

Bradley and Lang proposed a model of estimating arousal[7] and valence[8], allowing

[7]Arousal refers to the physiological and psychological state of being stimulated to a point of perception.

[8]Valence represents the intrinsic attractiveness or averseness of an event, object, or situation.

for interpreting emotions from facial expressions (Bradley et al. 1992). In this model, emotions are depicted along continuous scales of arousal and valence, facilitating a dynamic representation of facial expression changes over time.

Rudovic et al. (Rudovic, Lee, Dai, Schuller & Picard 2018) further explored the heterogeneity in facial expressions, body movements, and physiological responses from learners as they watched the same short video. They employed high-dimensional data embedding techniques to project these features into a two-dimensional space, revealing distinct distributions for different children's data. Such heterogeneity poses significant challenges to the accuracy of ML algorithms in classifying emotions across different children. The diverse expressions of emotions can lead to reduced classification accuracy when standard ML models are applied without adjustments for individual differences. This highlights the necessity for personalized approaches in the automatic recognition of emotional states, particularly in educational settings where understanding and addressing individual emotional responses can greatly enhance learning outcomes.

Rudovic, Utsumi, Lee, Hernandez, Ferrer, Schuller & Picard (2018) proposed CultureNet which is a deep neural network architecture specifically designed to account for cultural differences in behavioral data. Their work compares a 5-layer neural network, trained on a mixed dataset of Asian and European children, with CultureNet, which consists of two independent sub-networks, each trained exclusively on data from one cultural group. The study evaluates model performance under different conditions: when both training and validation data share the same cultural background versus when they are culturally mixed, and across the entire dataset. CultureNet shows superior performance compared to the baseline models. This approach reveals a key insight: children display culturally distinctive behaviors, expressing the same emotion through different facial expressions and body gestures, which can lead to significant estimation errors in cross-cultural evaluations. The findings suggest that incorporating culture-specific models can improve classification accuracy. However, the improvement in accuracy values is relatively modest, which raises questions about the practical significance of these improvements.

Building on previous work in Rudovic, Utsumi, Lee, Hernandez, Ferrer, Schuller & Picard (2018), researchers introduced the Personalized Perception of Affect Network (PPA-net) to classify affective states, namely arousal, valence, and engagement, in children from diverse cultural backgrounds. This network integrates multiple modali-

ties, including video recordings that capture facial expressions, head movements, body movements, poses, and gestures; audio recordings; and physiological signals such as heart rate, electrodermal activity, and body temperature. The architecture comprises six layers (Hahnloser et al. 2000). Layers 0 and 1 are responsible for pre-processing the multi-modal data, while layer 2 focuses on learning culture-specific features. Layer 3 is dedicated to gender-based differences, and layer 4 captures individual-specific nuances; finally, layer 5 generates the output predictions. The training strategy unfolds in three phases. In the first phase, all layers are trained on a combined dataset from all children, establishing a universal foundation. The second phase involves generating four separate neural networks tailored for Asian males, Asian females, European males, and European females. During this stage, layers 0 and 2 are copied from the global model and fixed, whereas layers 3 and 4 are fine-tuned using subgroup-specific data. In the third phase, individual networks are created by copying layers 0 through 3 from the appropriate subgroup network, and then layer 5 (and an additional layer, if applicable) is updated using each child's unique data. Experimental results indicate that PPA-net outperforms the baseline network. For instance, in the case of valence classification for Asian males, PPA-net achieves a mean accuracy of 0.59 ± 0.12, compared to 0.58 ± 0.12 for the baseline network. Although the improvement is modest (PPA-net's accuracy is within the traditional network's mean and standard deviation), they concluded that incorporating cultural, gender, and individual-specific adaptations can enhance the accuracy of affect classification, providing a nuanced approach that builds on and refines earlier models.

Shi et al. Shi et al. (2021) developed a system to better recognize children's emotional states, focusing on how excited or calm (arousal) and how positive or negative (valence) they feel. Their system combines different types of information, such as body movements, facial expressions, voice, and how children perform in games. Unlike earlier work by Rudovic et al. (Rudovic, Utsumi, Lee, Hernandez, Ferrer, Schuller & Picard 2018), their approach adjusts the model to work better for each individual child by using both general data from many children and specific data from each child. This helps the system provide more accurate results. For example, for one child, their method correctly identified emotional states 91% of the time, slightly better than using only general or individual data alone. However, while the accuracy is high, the improvement over simpler methods is relatively small.

Kort et al. (Kort et al. 2001) developed a comprehensive four-quadrant model

that underpins the MIT group's affective learning companion. This model conceptualizes effect along the two fundamental dimensions of arousal and valence, thereby dividing the affective space into four distinct regions. By continuously monitoring key facial features, such as the movements of the eyebrows, eyes, and mouth, the system maps subtle facial cues onto these quadrants to infer a learner's emotional state in real-time. The fully automated program leverages pattern recognition and machine learning techniques to achieve reliable classification, enabling the learning companion to adapt its instructional strategies based on the detected affective state. This integration of psychological theory with advanced computational methods exemplifies how automated systems can enhance learning by being sensitive to students' emotional cues.

In a related effort, D'Mello & Graesser (2013) introduced a dynamic decision network designed to interpret facial data and detect students' cognitive states during learning. Unlike static classifiers, this network incorporates temporal dynamics by continuously updating its predictions as new facial data is received. It not only captures instantaneous expressions but also tracks the evolution of these expressions over time, thereby distinguishing among cognitive states such as confusion, engagement, or boredom with greater accuracy. The dynamic decision network integrates both static facial features and their temporal changes, employing probabilistic models that balance immediate cues against longer-term patterns. This nuanced approach allows for a more context-aware and adaptive response, highlighting the potential of dynamic models to advance the design of intelligent tutoring systems that respond in real-time to the evolving cognitive and emotional signals from learners.

The integration of deep learning advances has significantly enhanced the field of facial expression recognition, particularly in identifying expressions related to MMP, such as confusion, frustration, surprise, and boredom. These prior developments provide a technical foundation for this research to interpret young pupils' facial expressions during learning activities.

2.7 Summary

This chapter began by highlighting the growing integration of AI techniques in educational platforms, which creates new opportunities to personalize and tailor educational support based on interpreting learners' responses. In the context of mathematical

learning, both standard and computer-based interventions are reviewed, illustrating their respective strengths and limitations. The chapter then focused on metacognitive monitoring, a central theme of this thesis, explaining the importance of MMP in mathematics learning and outlining the conventional approach to its estimation.

Recognizing limitations of conventional MMP estimation approaches, which often rely on self-assessment methods, are limited in their suitability for young pupils. This PhD research is motivated to investigate alternative solutions. Prior studies have demonstrated the relationship between facial cues and MMP in undergraduate students, suggesting the potential of non-verbal indicators for metacognitive monitoring in pupils. Building on this foundation, and supported by advances in deep learning that enable automatic and objective interpretation of facial cues, this research investigates how pupils' facial cues can be effectively used for MMP estimation. Furthermore, it explores the educational impact of interventions informed by this approach, aiming to enhance learning outcomes through timely and individualized support.

Chapter 3

User Study 1: Collecting Facial Cues in Metacognitive Monitoring of Pupils

---

When one door closes, another opens; but we often look so long and regretfully upon the closed door that we do not see the one which has opened for us.

— Alexander Graham Bell

This chapter reports research used, in part, to address RQ1 in Chapter 1. It reports the development of a smart game to stimulate children to perform metacognitive monitoring and reflect upon their past performance. The smart game was used to collect data for this PhD research. Content from this chapter was presented at the ACM IDC 2023 paper titled Supporting Children's Metacognition with a Facial Emotion Recognition-based Intelligent Tutor System (Ruan, X., Palansuriya, C., Constantin, A., & Tsiakas, K., 2023, June).

Metacognitive monitoring refers to the process of monitoring one's own thought processes and existing state of knowledge. As discussed in Chapter 2, research consistently demonstrates that children achieve better learning outcomes when they have better metacognitive monitoring performance (MMP). Interventions designed to support metacognitive monitoring are typically adapted based on post-task evaluations of MMP, such as confidence in judgments.

The work presented in this chapter was motivated to investigate the relationships between facial cues and MMP in pupils aged 7 to 11, building upon prior findings from undergraduate students discussed in Section 2.5. This investigation was conducted in

the context of pupils engaging in mathematics-related tasks within computer-based learning environments (CBLEs). Specifically, it presents the first user study, involving 168 pupils aged 7 to 11 from two provinces in China, where facial cues were recorded as participants engaged in cognitive tasks, including inhibitory control, working memory, and target switching, related to mathematics, physics, and other science subjects (Brookman-Byrne et al. 2018, Träff et al. 2019). The analysis identified a set of facial cues associated with MMP in pupils. Overall, the findings underscore the potential of using facial cues for real-time, adaptive assessment of pupils' metacognitive monitoring.

3.1 Introduction to the User Study 1

As discussed in Section 2.3 and Section 2.4, metacognitive monitoring plays a pivotal role in self-regulated learning (SRL). In learning, metacognitive monitoring interventions are provided to enable learners to accurately reflect on what they know and what they do not know (Zumbach et al. 2020).

Affective states refer to the internal experiences associated with feelings, emotions, and moods. Understanding the affective states experienced by learners during self-regulated learning (SRL) is essential, as these emotional responses can signal moments of cognitive dissonance or the recognition of knowledge gaps, both of which play a critical role in reflecting effective metacognitive processes (Mandler 1989). In Chapter 2, prior studies about the interaction between affective states and metacognitive monitoring have been reviewed (Mandler 1989, Efklides 2011). In addition, some of the research also observed that emotions such as surprise, frustration, and confusion often emerge when discrepancies occur (Taub & Azevedo 2018, Cloude et al. 2020, Taub et al. 2021). Building on these reviewed findings in undergraduate students, Chapter 2 highlighted the potential of estimating pupils' MMP through interpreting their emotional responses. Thus, this chapter deepens the investigation of how facial cues, such as facial expressions, head gestures, and gaze directions, can serve as indicators of pupils' MMP.

To achieve this, the chapter focuses on a user study that was designed and conducted using a tailored serious game, facilitating the collection of comprehensive data on pupils' facial cues and MMP data. This contribution establishes a foundation for estimating MMP based on nuanced observations of affective states.

The chapter begins with a review of related work that informed the design of the user study, detailed in Section 3.3. The research questions are presented in Section 3.2. The application developed specifically for the user study is introduced in Section 3.4. The procedure of the user study is outlined in Section 3.5, followed by a description of the data pre-processing methods in Section 3.6. The results of the user study are then presented, along with a discussion of the findings in Section 3.8 and Section 3.9.

3.2 Research Questions

Given the specific population (pupils aged 7 to 11) and the deployment scenario (maths in computer-based learning), two specific sub-questions, RQ1.1 and RQ1.2, were investigated to address RQ1 from Chapter 1. These sub-questions were designed to extend prior evidence on the relationship between MMP and task performance in young learners while they solve mathematics-related tasks in CBLEs, and to further explore the correlation between MMP and facial cues in children.

3.3 Related Works

Chapter 2 discussed the correlation between affective states/emotions and MMP, so this section highlights related work about collecting facial data and potential factors that need to be controlled.

3.3.1 Development of Datasets from Collecting Facial Data in Metacognitive Monitoring

Previous studies aimed at building datasets from collecting facial data in educational settings have used a variety of stimuli, including engaging participants to read materials or solve predefined tasks. These stimuli are designed to replicate the affective states that participants typically experience in real-world educational environments (Linson et al. 2022, Gupta et al. 2016, Singh et al. 2023). Typically, such studies employed readily accessible devices, such as webcams, to capture subjects' facial cues. For instance, Linson et al. (Linson et al. 2022) focused on investigating real-time feedback on student engagement in the remote-learning scenario. They invited students

to watch a pre-recorded linear algebra course and recorded facial behaviors through laptop cameras. The duration of the course recording is limited to 10 minutes to optimize the balance between comprehensive data capture and considerations like file size and attention span. Similarly, Gupta et al. (Gupta et al. 2016) aimed to capture engagement 'in the wild', thus bridging the gap between controlled laboratory settings and real-world applicability in domains such as online learning and healthcare. Their setup involved high-resolution video recordings via a full HD webcam, allowing for detailed observations of facial cues and better insights into user engagement.

A common method used to assess MMP is Judgment of Learning (JOL) questions (Taub et al. 2021), where participants are requested to predict their ability to recall previously studied material. The Judgment of Confidence (JOC) questions (Maras et al. 2019) are also frequently employed, allowing researchers to compare individuals' perceived certainty with their actual performance. JOC questions are particularly popular with children, as they are more straightforward for children to assess their achieved performance rather than predicting future outcomes (Mandler 1989). Inspired by these methodologies, the user study aimed to develop a data collection prototype that engages children in metacognitive monitoring by having them answer JOC questions after completing tasks.

3.3.2 Factors Affecting Facial Cues in Learning Environments

Facial cues serve as critical nonverbal indicators in learning environments, reflecting the intricate interplay of cognitive, emotional, and environmental factors. This section reviews essential factors that must be controlled in the user study to accurately capture facial cues associated with metacognitive monitoring. Identifying and managing these variables is necessary to ensure the reliability and validity of the resulting observations and findings.

3.3.2.1 Cognitive Load

Cognitive load significantly influences facial cues as learners engage with complex materials. Research by D'Mello et al. (D'Mello et al. 2007) demonstrates that as cognitive load increases, learners may exhibit expressions of confusion, frustration, and boredom. Furthermore, the design of the user interface (UI) can also exacerbate cognitive load during user interactions with an application, as noted by Oviatt (Oviatt

2006). When the UI presents information that overwhelms users, it often leads to increased signs of boredom and disengagement. This evidence highlights the critical need to manage cognitive load in the user study. It is essential to ensure that the cognitive load borne by subjects is solely attributable to the task demands and not confounded by external factors such as poor user interface design.

3.3.2.2 Learning Environment

The learning environment, encompassing everything from classroom design to interpersonal dynamics, plays a significant role in shaping how emotions are expressed and perceived. Supportive and engaging environments are known to foster a broader spectrum of emotional expressions (Pekrun & Linnenbrink-Garcia 2012, Ainley 2006). Therefore, it is crucial to create an environment that is not only conducive but also motivating to subjects, as this is vital for gathering meaningful data on facial cues.

Recognizing the importance of minimizing the influence of these diverse external factors on pupils, the application used in this user study was designed to reduce these factors' impact on facial cues, ensuring that the cues primarily reflected MMP.

In summary, this section reviewed the literature on facial data collection in educational settings and highlighted critical factors for ensuring data quality and reliability. Building on these insights, the remainder of this chapter presents the design of a data collection tool, the development of stimuli to effectively elicit metacognitive monitoring, and the analysis of the collected data to address the outlined research questions.

3.4 The Meta-Brainhood Prototype Application

To collect facial data from pupils, a smart game, Meta-Brainhood, was used. The game was built on Brainhood, a previously developed educational game for children (Tsiakas et al. 2020).

3.4.1 Brainhood

Interaction design for children is a specialized field within user experience research that focuses on adapting UI and interactive experiences to align with the cognitive, emotional, and physical capabilities of children. To support this work, a collaboration

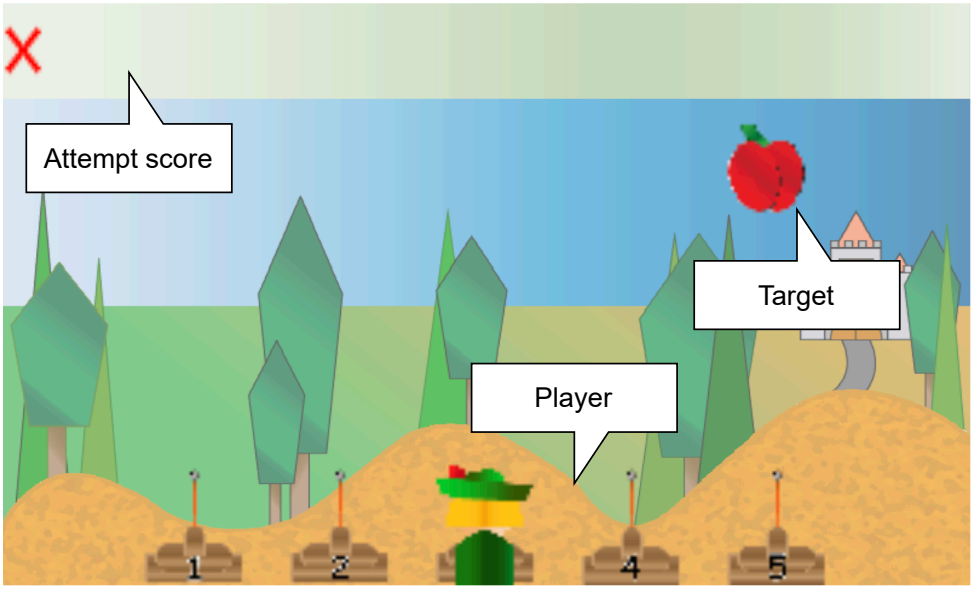


Figure 3.1: Brainhood game environment

was established with Dr. Konstantinos Tsiakas and Prof. Panos Markopoulos, drawing on their expertise and the Brainhood interface they developed (Tsiakas et al. 2020).

Brainhood is a cognitive training serious game that challenges children to control an archer character and hit targets appearing in various positions on-screen, see Figure 3.1. The Brainhood platform was selected because it is specifically designed to enhance generic cognitive skills crucial for solving mathematics problems, as supported by literature (Fanari et al. 2019, Brookman-Byrne et al. 2018, Li et al. 2020). The game's design is modular, based on several variables: the number of target/player positions, target types, player colors, and task rules. Players encounter two target types (red/green apples, left/right birds), two player colors (green, red), and seven possible target/player positions. This parametric design allows for a diverse range of game configurations to accommodate different player skills and preferences.

The prototype of Brainhood features three primary task parameters: rules, difficulty, and speed. The rules segment includes four types: Move and Shoot, Avoid Birds, Remember Targets, and Switch Targets, each introducing different goals and complexities. These rules can be combined in 15 different ways, offering varied gameplay complexity and difficulty. Difficulty levels adjust the number of target locations (3, 5, 7), while speed settings control the frequency of target appearances (Fast at 1 second per target, Medium at 1.5 seconds, Slow at 2 seconds). Altogether, these

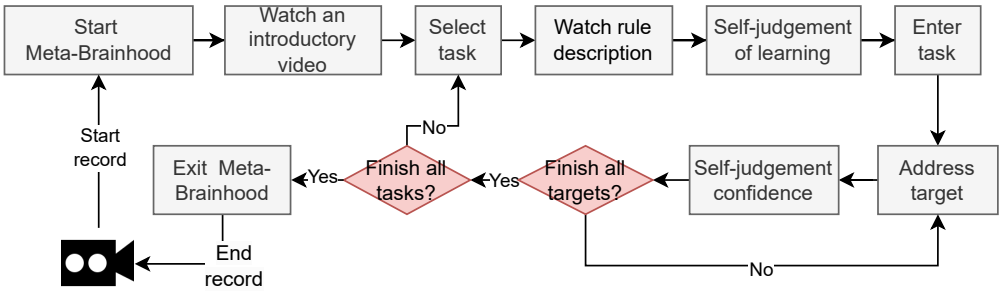


Figure 3.2: The workflow in Meta-Brainhood

parameters result in 195 potential task configurations, each designed to test and enhance specific cognitive abilities.

3.4.2 New Variant from Brainhood - Meta-Brainhood

The Brainhood prototype (Tsiakas et al. 2020) was enhanced to develop Meta-Brainhood, which is a modified version tailored for pupils aged 7–11 in China. This platform integrates Brainhood tasks designed for self-regulated game-based learning and includes functionality for collecting MMP data and facial cues data. An overview of the workflow of Meta-Brainhood is presented in Figure 3.2. Notably, Meta-Brainhood was specifically designed to simulate various cognitive tasks related to mathematical learning and to capture pupils' facial responses corresponding to those tasks.

Affective states are indicated by facial cues, such as expressions, head gestures, and gaze directions (Ekman & Friesen 1978). Ekman et al.'s work showed that affective states like happiness, sadness, anger, fear, surprise, and disgust are distinctly encoded in facial cues (Ekman 1999). Beyond expressions, head gestures and gaze directions both provide additional insights into an individual's affective state. According to Argyle et al., gaze direction not only indicates where attention is focused but also reflects levels of interest and the dynamics of social interactions, important factors in collaborative learning environments (Argyle et al. 1994). Accordingly, new functionalities were implemented in Meta-Brainhood to enable the collection of facial cues data.

Regarding factors that potentially affect facial cues in the user study. To reduce cognitive load, the platform was adapted to be more accessible for younger users.

(a) The introduction video at the beginning of Meta-Brainhood. 1: the content of the introduction video; 2: the progress bar of the introduction video.

(b) Welcome page of Meta-Brainhood. 1: 'Hello, welcome to Brainhood!'; 2: Six cognitive tasks; 3: Test start button; 4: The button to review awarded coins; 5: Test progress.

Figure 3.3: Modifications of the original version of Meta-Brainhood.

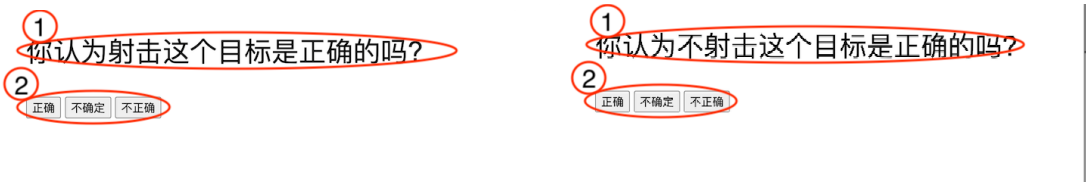
Table 3.1: The six tasks in the Meta-Brainhood and the corresponding cognitive tasks. (IC: inhibitory control; VSWM: visual and spatial working memory; TS: target switching.)

Task ID	**Task name**	**Cognitive tasks**
Task 1	Avoid birds	IC
Task 2	Remember targets	VSWM
Task 3	Target switch	TS
Task 4	Avoid birds & Remember targets	IC & VSWM
Task 5	Target switch & Remember targets	TS & VSWM
Task 6	Avoid birds & Remember targets & Target switch	IC & VSWM & TS

A brief instructional video is inserted as the beginning of the Meta-Brainhood, as shown in Figure 3.3a. This video provides pupils with a clear demonstration of the game objectives, component descriptions, overall flow, and key game-play settings. Additionally, the task selection page was simplified by presenting only six tasks, as highlighted in the red circle 2 in Figure 3.3b. This approach minimizes cognitive load by streamlining the information available at the game's outset. Detailed descriptions of the six cognitive tasks involved are listed in Table 3.1. They consist of different combinations of cognitive tasks. The first three tasks include three basic cognitive skills: inhibitory control (IC) (Brookman-Byrne et al. 2018), visual and spatial working memory (VSWM) (Fanari et al. 2019), and target-switch (TS) (Li et al. 2020). The rest of the three tasks combine tasks from these basic ones, see Table 3.1 (Tsiakas et al. 2020, Ruan et al. 2023). These cognitive skills have been shown to play a critical role in mathematics, physics, and other science subjects (Brookman-Byrne et al. 2018, Träff et al. 2019).

To enhance pupil engagement and foster a comfortable learning environment, the existing reward system was modified for the user study. Specifically, the traditional 'scores' were replaced with 'coins', as highlighted by the red ellipse 4 in Figure 3.3b. This adjustment enabled pupils to earn coins based on their performance, which could subsequently be exchanged for gifts. Furthermore, to ensure a familiar and conducive learning atmosphere, the user study was conducted in the pupils' regular classroom setting.

To collect pupils' MMP data, a JOC question was implemented, requiring pupils



(a) The JOC question. 1: Was it right to shoot that target? 2 (from left to right): It is correct, Not sure, It is wrong.

(b) The JOC question. 1: Was it right to miss that target? 2 (from left to right): It is correct, Not sure, It is wrong.

Figure 3.4: The JOC question used in Meta-Brainhood.

to self-assess their confidence in their responses following each attempt (see the self-judgment of confidence in Figure 3.4). The system poses the questions: 'Was it right to shoot that target?' and 'Was it right to miss that target?'. Pupils rate their confidence using a scale that includes the options: 'It was correct', 'Not sure', and 'It was wrong'.

The Meta-Brainhood interface underwent formative evaluation by three experts in child–computer interaction. The experts explored the application and participated in a semi-structured interview focused on the system's ease of use and its appropriateness for children. Based on their feedback, the interface was refined. For example, labels were modified to enhance clarity for children, and language was simplified for younger users. Overall, the experts concluded that the interface is both easy to use and suitable for children.

3.5 User Study Procedure

This section outlines the recruitment process for involving pupils in the user study and details the procedures for data collection, transfer, and storage. The research was conducted with a strong understanding of children's rights and adhered strictly to ethical approval protocols, ensuring that all procedures respected the rights and well-being of the child participants. Ethical approval for this study was obtained from the Ethics Committee of the University of Edinburgh, School of Informatics (Approval Code: RT#6963) in November 2023, and from the Ethics Committee of Shandong Normal University (Approval Code: SDNU2023050) in October 2023. The related ethics application documents are available in (github/affect2mmp 2025).

Figure 3.5: Location map of the user study 1

3.5.1 Recruitment and Participation of Children

The user study was conducted in mid-November 2023 in China. We advertised this study in target schools. In the end, we involved 186 pupils recruited from two primary schools.

Prior to the study, information sheets and participant consent forms[1] were distributed, including versions for both children and their parents. Participation in the study was voluntary, allowing pupils to schedule their time for the workshop and providing the flexibility to interrupt or withdraw at any point. Signed consent forms and information sheets from both pupils and their parents were obtained for participation and data collection.

Data from 14 pupils were excluded due to interruptions during the workshop, and data from 4 pupils were excluded due to technical issues with webcam activation. The final dataset used for analysis comprised 168 pupils, with a mean age of 9.83 years and a standard deviation of 1.62, including 80 males and 88 females. The locations of two schools are marked in Figure 3.5.

3.5.2 Materials

Meta-Brainhood was installed on laptops for this study. All content was translated from English into Chinese to ensure accessibility in the pupils' native language. Ses-

[1] A copy version of these documents is available on the GitHub site, (github/affect2mmp 2025).

sions were conducted in classrooms at each participating primary school to provide a familiar and comfortable environment. Each classroom was equipped with desks, chairs, power chargers, laptops, and a whiteboard.

3.5.3 Procedure

The study was conducted over a period of seven days in two schools simultaneously in Region A and B, with six sessions held per day. Each pupil participated in only one session, which was scheduled according to a pre-arranged timetable. In each session, pupils worked individually on laptops and could ask for help by raising their hands. Before beginning, laptop webcams were set up for the pupils, and a briefing was provided on completing six tasks within Meta-Brainhood. Their objective was to accumulate as many reward coins as possible. After inputting basic information (age, gender, nationality), the application played an introductory video outlining how to engage with the game, as shown in Figure 3.3a.

Pupils spent approximately 35 minutes completing all tasks. After finishing, participants had the option to exchange their earned coins for gifts (toys, pens, bags, and notebooks) and decide whether to proceed with another round or conclude their participation in the workshop.

3.5.4 Data Collection

As for pupils' MMP, the JOC questionnaire automatically appeared on the screen after each attempt. In total, pupils were asked to make 60 attempts and respond to 60 JOC. Then Meta-Brainhood automatically stores pupils' scores and JOC responses in a document on the laptop.

To record facial cues, the Meta-Brainhood recorded pupils' facial area through the webcam included in the laptop. Based on the timestamps of self-rated JOCs, the Meta-Brainhood clipped periods during which children were reporting JOC (that is, from the moment the JOC questionnaire appeared on the screen to when the children submitted their responses and the questionnaire closed). Consequently, each child contributed 60 video clips of performing metacognitive monitoring, and the video clips were stored on the laptop's built-in hard disk.

3.6 Data Acquisition and Pre-processing

This section introduces the measurements of MMP, performance, and facial cues collected from the user study. Facial cues include facial expressions, direction of the gaze, and head gestures. All of these measurements will be used to address the research questions in this chapter.

3.6.1 Measurements for MMP and Cognitive Task Performance in Meta-Brainhood

It was first important to develop the measurements used to assess MMP and cognitive task performance among pupils in the user study.

Attempt score (AS): In task $t \in \{1, 2, \ldots 6\}$, pupils are given 10 attempts ($p \in \{1, \ldots 10\}$) and are requested to shoot or miss a target. The $AS_{t,p}$ is 1 if the attempt p is correct (shot or miss) in task t, or 0 if pupils incorrectly shoot or miss the attempt.

$$AS_{t,p} = \begin{cases} 1, & \text{if pupils correctly address the attempt,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

Task score (TS): The TS_t is calculated by:

$$TS_t = \sum_{p \in [1,\ldots 10]} \{AS_{t,p}\}. \tag{3.2}$$

Attempt confidence rating (ACR): Pupils' JOC responses for attempt p of task t are valued on a scale from 0 to 1:

$$ACR_{t,p} = \begin{cases} 0, & \text{if respond 'It is wrong',} \\ 0.5, & \text{if respond 'Not sure',} \\ 1, & \text{if respond 'It is correct'.} \end{cases} \tag{3.3}$$

Attempt Calibration Index (ACI): To measure pupils' MMP for attempt p of task t, the approach outlined by (Schraw 2009) was followed. The $ACI_{t,p}$ was calculated as the squared deviation between $ACR_{t,p}$ and $AS_{t,p}$, as shown in Equation 3.4. This formula computes the squared deviation between confidence ratings and actual performance, where higher values (closer to 1) indicate higher MMP.

$$ACI_{t,p} = 1 - (ACR_{t,p} - AS_{t,p})^2 \tag{3.4}$$

Task Calibration Index (TCI): For the MMP of each task, the TCI_t is calculated by:

$$TCI_t = \frac{1}{10} \sum_{p \in [1,...10]} ACI_{t,p}. \tag{3.5}$$

3.6.2 Measurements and Pre-processing for Facial Cues in Addressing JOC Question

OpenFace (Baltrusaitis et al. 2018) is a comprehensive, anatomically based system for describing all visually discernible facial cues. In this study, OpenFace was adopted to numerically measure pupils' facial cues, including Action Units (AUs), gaze directions, and head gestures.

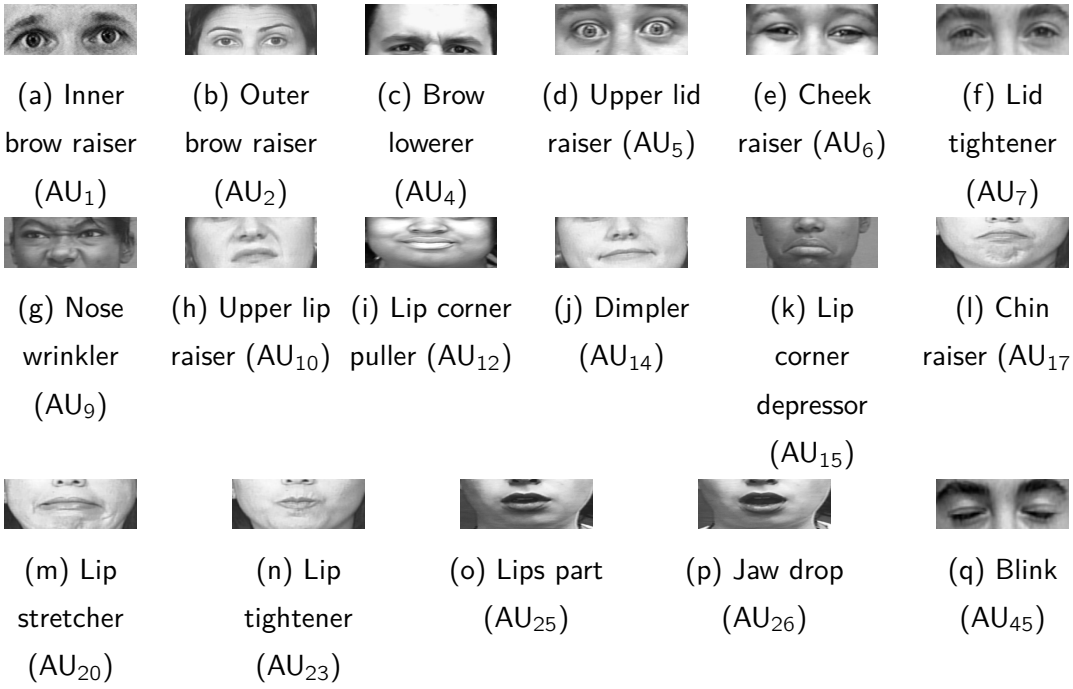
Action Units: These represent individual components of muscle movement, called AUs, based on the Facial Action Coding System (Ekman & Friesen 1978). OpenFace extracted the facial AUs from face images coming from the facial recording and valued each AU from 0 (weak) to 5 (strong). All 17 AUs[2] detected by OpenFace are displayed in Figure 3.6.

Head pose: Head pose is represented by six values, such as, location and rotation. The location of the head is the relative location to the camera in millimeters (positive Z is away from the camera).

Head rotation: In OpenFace, head's rotations are represented along three axes: X, Y, and Z. Rotation along the X-axis, known as pitch, corresponds to the head moving up and down where a positive pitch means the person is looking upward, while a negative pitch indicates they are looking downward. Rotation along the Y-axis is referred to as yaw and reflects left-to-right movement, a positive yaw means the head is turned to the right, and a negative yaw means it is turned to the left. Lastly, rotation along the Z-axis is called roll, describing the head tilting from side to side, a positive roll tilts the head toward the left shoulder, whereas a negative roll tilts it toward the right shoulder.

Gaze direction: Gaze direction from OpenFace is averaged for both eyes and represented by two values of radians in world coordinates. For example, if a person is looking left-right (horizontal) this will result in a change of gaze direction (the first value) from positive to negative, and, if a person is looking up-down (vertical) this

[2]AU numbers are not sequential because some were reserved for muscle groups that were later split, merged, or discarded.



(a) Inner brow raiser (AU_1) (b) Outer brow raiser (AU_2) (c) Brow lowerer (AU_4) (d) Upper lid raiser (AU_5) (e) Cheek raiser (AU_6) (f) Lid tightener (AU_7)

(g) Nose wrinkler (AU_9) (h) Upper lip raiser (AU_{10}) (i) Lip corner puller (AU_{12}) (j) Dimpler (AU_{14}) (k) Lip corner depressor (AU_{15}) (l) Chin raiser (AU_{17})

(m) Lip stretcher (AU_{20}) (n) Lip tightener (AU_{23}) (o) Lips part (AU_{25}) (p) Jaw drop (AU_{26}) (q) Blink (AU_{45})

Figure 3.6: Action units detected through OpenFace in this PhD thesis. Facial images from Ekman et al. (2002). Note: The 17 AUs shown are those supported by OpenFace. These represent the AUs OpenFace supports with reliable detection accuracy, focusing on those most commonly associated with expressive facial behavior.

will result in a change of gaze direction (the second value) from negative to positive, if a person is looking straight ahead both of the angles will be close to 0.

The facial recordings input to OpenFace contain 30 frames per second. OpenFace was used to extract values for facial cues (AUs, head gestures, and gaze directions) from each frame. As a result, the raw output file includes 30 instances of each facial cue per second. The raw facial cue data corresponding to each individual's metacognitive monitoring segments (comprising a total of L frames) were pre-processed to reduce noise and improve data reliability. These pre-processing steps included removing missing values, standardizing the scale, and smoothing the values, as validated in Taub et al. (2021).

1. Removing missing values: For each facial cue, missing values were replaced with the corresponding average value within the same video.

2. Standardize scale: The rescaling process was applied to eliminate influences from the subject difference by converting the intensity value of each AU to the

Figure 3.7: Judgment of learning questionnaire. 1: How well do you understand the rule? 2: Very bad, bad, not bad, good, very good (left to right). 3: Submit.

range of $[0-1]$. The formula used for rescaling is:

$$\text{rescaled}(x) = \frac{x - min(x)}{max(x) - min(x)}, \tag{3.6}$$

x is the specific feature value of facial cues, and the $min(x)$ and $max(x)$ represent the minimum value and maximum value of the corresponding feature in the recording.

3. Smoothing of facial cues: To reduce noise and smooth values of facial cues of L frames, a mean value filter with a window size of 11 was applied to each rescaled facial cue. For example, the value of frame l was replaced by the average of itself and its five preceding and five succeeding frames, see Equation 3.7.[3]

$$\text{filtered}(x_l) = \frac{1}{11} \sum_{i=-5}^{5} x_{l+i}, \quad 5 < l \leq L - 5 \tag{3.7}$$

3.7 Statistical Summary of Pupils' Performance and Facial Cues

This section presents the relevant data collected from pupils in the user study, as outlined in Section 3.6. The following descriptions cover pupils' preparation of the rules prior to completing tasks, task scores, and MMP during task performance.

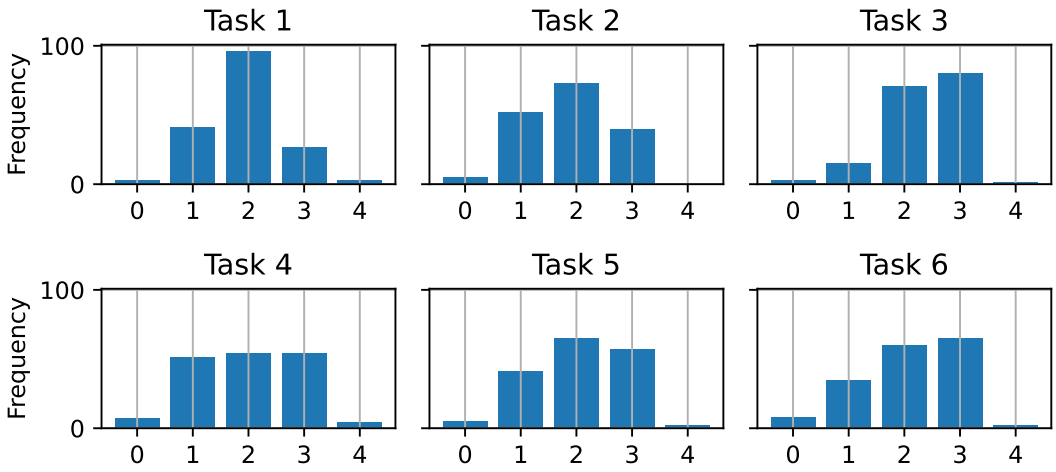


Figure 3.8: The distribution of pupils' Judgment of Learning responses in tasks.

3.7.1 Pupils' Preparation on Tasks

Before each of the six tasks, the rules were presented through a video demonstration, allowing pupils to replay or skip sections according to their individual learning pace. This approach aimed to ensure an adequate level of preparation prior to beginning each task. Pupils' confidence in their understanding of the rules was then assessed using a JOL questionnaire, with scores ranging from 0 (very bad) to 4 (very good), see Figure 3.7.

For Task 1 (IC test), pupils reported a medium level of confidence ('Not bad'), with an average JOL score of 1.92, indicating that they felt somewhat comfortable and had an adequate understanding of Task 1.

For Task 2 (VSWM test), confidence levels showed a slight decrease, with an average JOL score of 1.87. Despite this small drop, the pupils still felt relatively confident ('Not bad') about their understanding of Task 2.

For Task 3 (TS test), this task saw an increase in confidence, with an average JOL score of 2.36, above the 'Not bad' confidence level. This suggests that pupils felt 'Not bad' about Task 3, and the rule complexity in this task may have been more manageable for pupils.

For Task 4 (IC&VSWM test), confidence dips in this task, with an average JOL score of 1.98, suggesting that the pupils felt 'Not bad' in their understanding of Task

[3]This smoothing process was essential to mitigate the effects of transient, non-informative facial movements (e.g., minor twitches or brief lighting changes) that do not reflect genuine metacognitive monitoring behaviors but could otherwise introduce noise into the analysis (Taub et al. 2021).

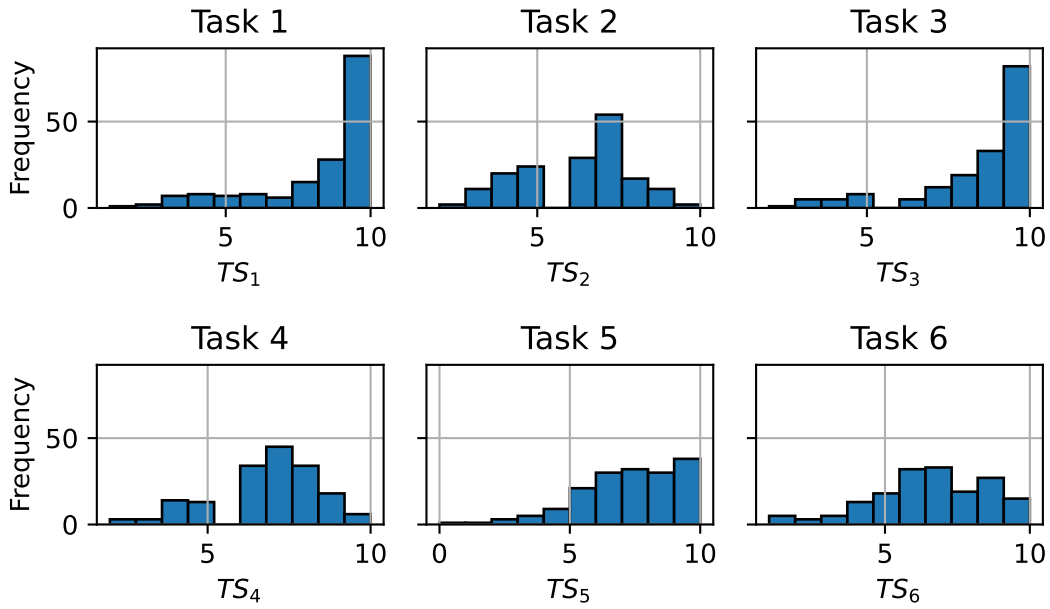


Figure 3.9: The distribution of pupils' task scores in tasks.

4.

For Task 5 (TS&VSWM test), the confidence levels remained around the 'Not bad' level, with an average JOL score of 2.06, suggesting that the pupils felt 'Not bad' in their understanding of Task 5.

For Task 6 (IC&TS&VSWM test), pupils again reported a 'Not bad' confidence level, with an average JOL score of 2.11. This value suggests that pupils felt moderately confident in their performance on the task.

3.7.2 Pupils' Task Scores

The distribution of pupils' scores (TS) across six different tasks is presented in the histograms shown in Figure 3.9. These distributions suggest varying levels of task difficulty, reflecting differences in the cognitive demands placed on the pupils.

Pupils' average TS in Task 1 and Task 3 are 8.44 and 8.59, respectively. The distributions show a concentration of higher scores, implying that these tasks were relatively easier for most pupils, which may be due to the nature of the cognitive skills tested.

Pupils' average TS in Tasks 2, 4, 5, and 6 are 6.14, 6.76, 6.82, and 6.68, respectively. These distributions exhibit a broader spread of scores. Considering these tasks

are based on the visuospatial working memory (VSWM) cognitive test (refer to Table 3.1), this pattern suggests that VSWM represents a more challenging cognitive skill for pupils aged 7 to 11, as indicated by the more moderate and dispersed scoring.

It is clearly identified that there is a gap in the task score distribution of Tasks 2, 3, and 4, indicating a possible bimodal trend. This suggests that pupils tended to either perform well or poorly on these tasks, with fewer falling in the mid-range, which may reflect differences in understanding, prior knowledge, or strategy use for those tasks.

The score distributions for Tasks 2, 4, 5, and 6 indicate a clear educational need to enhance VSWM among pupils in this age group. Given the critical role that VSWM plays in learning processes such as understanding spatial relationships and managing visual information in mathematics, these findings were communicated to the principal of the school where the study was conducted.

3.7.3 Pupils' MMP in Tasks

Participants were asked to self-reflect on their task attempts and report their confidence using the JOC following each attempt. The distribution of pupils' MMP, measured by TCI, across six different tasks is presented in Figure 3.10.

Task 1 and Task 3 predominantly show MMP clustering at higher values. Pupils' average TCI in those tasks is 0.78 and 0.79, respectively. This suggests that most pupils either selected 'It is correct' when making the correct attempt or selected 'It is wrong' when making the incorrect attempt.

Tasks 2, 4, 5, and 6 show more uniformly spread distributions, though with a slight skew towards lower MMP. The average TCI in these tasks is 0.58, 0.59, 0.59, and 0.58, respectively. This pattern (less than 0.75) shows that most pupils don't know whether their attempts were correct or incorrect. This could indicate a consistent challenge across the cohort, suggesting that these tasks engaged cognitive skills that are challenging for this age group.

Combining distributions as depicted in Table 3.2, it offers valuable insights into the impact of pupils' MMP on task scores. Notably, the similar distribution in those figures suggests that pupils tend to achieve higher performance in tasks where they have a clear understanding of their own capabilities. This observation underscores the importance of MMP in learning, as students who effectively evaluate their knowledge

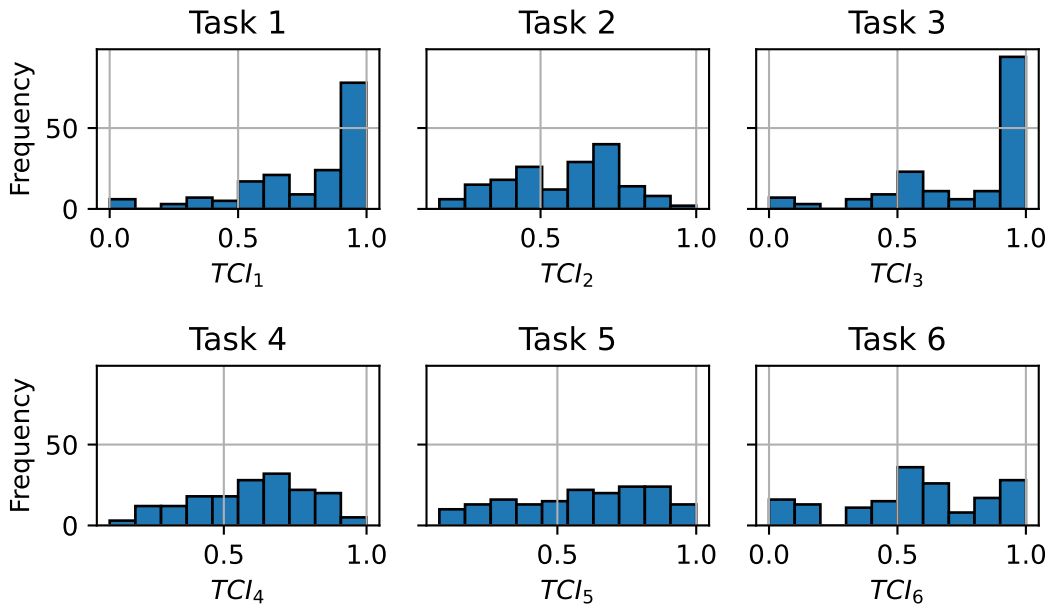


Figure 3.10: The distribution of pupils' MMP (represented by Task Calibration Index (TCI)) in tasks.

Table 3.2: Summary of Judgment of Learning, Task Scores, and Metacognitive Monitoring Performance by task

Task ID	Avg. JOL	Avg. TS	Avg. TCI
Task 1	1.92	8.44	0.78
Task 2	1.87	6.14	0.58
Task 3	2.36	8.59	0.79
Task 4	1.98	6.76	0.59
Task 5	2.06	6.82	0.59
Task 6	2.11	6.68	0.58

Table 3.3: Measuring the Pearson-coefficients between pupils' MMP and cognitive tests. IC: inhibitory control; VSWM: visual and spatial working memory; TS: target switching.

Task ID	Cognitive tests	Correlation by tasks	
		$\rho_{(TCI_t, TS_t)}$	p
Task 1	IC	0.67	< 0.01
Task 2	VSWM	0.42	< 0.01
Task 3	TS	0.55	< 0.01
Task 4	IC&VSWM	0.47	< 0.01
Task 5	TS&VSWM	0.47	< 0.01
Task 6	IC&TS&VSWM	0.42	< 0.01

and skills are better equipped to tackle tasks successfully. Such findings highlight potential areas for developing educational strategies to enhance metacognitive monitoring among students, thereby improving their overall task performance.

Based on these statistical descriptions, RQ1 is investigated in Section 3.8.1.

3.8 Experiment Results

Drawing on pupils' task scores, MMP, and measured facial cues, this section addresses the two research questions outlined at the beginning of the chapter. First, it examines the impact of pupils' MMP on their task performance, and then identifies the facial cues that show significant correlations with MMP.

3.8.1 RQ1.1

This section addresses RQ1 by reporting the impact of pupils' MMPs on their task scores in Meta-Brainhood. The Pearson correlation coefficient was calculated between pupils' calibration indexes (TCI_t) and their cognitive task scores (TS_t), as presented in Table 3.3. In addition, a linear regression analysis was conducted to examine the relationship between task scores and MMPs, with the results shown in Figure 3.11 and Table 3.4.

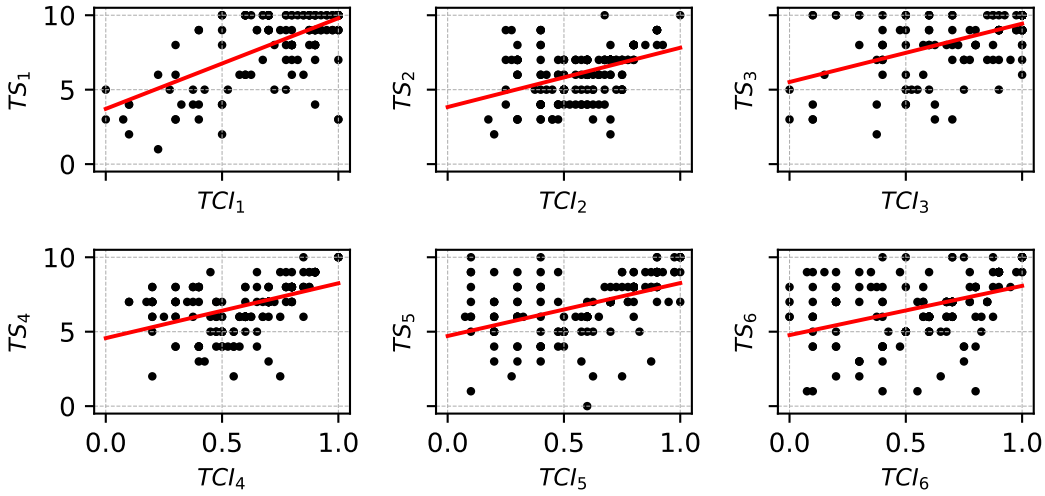


Figure 3.11: Linear regression between task scores and MMP. Note that the Y axis (TS) is task score, and the X axis (TCI) is MMP; see Equation 3.2 and Equation 3.5.

Based on the numerical results presented in Table 3.4 and the visual trends illustrated in Figure 3.11, the following observations can be drawn.

For Task 1 (IC test), there was a strong correlation ($r = 0.67$) between TCI and TS, indicating that higher MMP is strongly associated with better task performance. The regression analysis indicates that task scores increase by 6.09 points for each one-point increase in TCI, starting from a base score of 3.71.

For Task 2 (VSWM test), there was a moderate correlation ($r = 0.42$), suggesting a positive but less pronounced relationship between TCI and TS. From the regression analysis, each unit of increase in TCI is associated with an increase of 3.98 points in

Table 3.4: Linear Regression Results Between MMP and Task Scores

Task ID	**Correlation (r)**	**Slope**	**Intercept**	**Interpretation**
Task 1	0.67	6.09	3.71	Strong positive relationship
Task 2	0.42	3.98	3.83	Moderate positive relationship
Task 3	0.55	3.92	5.51	Moderate positive correlation
Task 4	0.47	3.68	4.57	Moderate positive correlation
Task 5	0.47	3.55	4.71	Moderate positive correlation
Task 6	0.42	3.31	4.77	Moderate positive correlation

task scores, with an intercept of 3.83.

For Task 3 (TS test), there was a moderately strong correlation (r = 0.55) with task scores. The regression analysis indicates that each one-point increase in TCI is associated with a 3.92-point increase in task score, starting from a baseline score of 5.51. This suggests that TCI makes a significant contribution to improved task performance.

For Task 4 (IC & VSWM test), there was a moderate correlation (r = 0.47). From the regression analysis, a one-unit increase in TCI leads to a 3.68-point increase in task scores, starting from a baseline score of 4.57.

For Task 5 (VSWM & TS test), there was also showed a moderate correlation (r = 0.47). The regression analysis shows that task scores increase by 3.55 points for each one-point increase in TCI, beginning from 4.71.

For Task 6 (IC & VSWM & TS test), the correlation value (r = 0.42) was reported, and it is similar to Task 2. The regression analysis shows that task scores are quantified as increasing by 3.31 points for each one-point rise in TCI, with an intercept of 4.77.

3.8.2 RQ1.2

This section investigates the relationships between pupils' MMP and their facial cues expressed during cognitive tasks. Following the data pre-processing procedures described in Section 3.6.2, numerical results were obtained for all measured facial cues from the metacognitive monitoring clips. The expression of each facial cue was represented using the average value and the standard deviation of its measurements.

Facial cues during metacognitive monitoring were analyzed by measuring the average intensity and standard deviation of expression levels. These measurements were then correlated with pupils' MMP, as detailed in Table 3.5 and Table 3.6.

Based on the illustrated significant correlated facial cues[4] in Table 3.5 (average values) and Table 3.6 (standard deviation values), the facial expressions that are significantly correlated with MMP, either at the average level or the standard deviation level, were identified. These include the outer brow raiser (AU_2), brow lowerer (AU_4), cheek raiser (AU_6), lid tightener (AU_7), lip corner puller (AU_{12}), dimpler (AU_{14}), lip corner depressor (AU_{15}), lip tightener (AU_{23}), lips part (AU_{25}), and blink (AU_{45}).

[4]Denoted with a symbol *.

Additionally, significant correlations were observed with gaze direction (horizontal G_x and vertical G_y), head location within the webcam frame ($H(loc.)_x$, $H(loc.)_y$) and $H(loc.)_z$), and head rotations (up and down ($H(rot.)_x$), left and right ($H(rot.)_y$)).

3.8.2.1 How Does Variation in MMP Affect the Expression of Identified Facial Cues During Tasks?

Cognitive tasks in Meta-Brainhood are designed to engage pupils in metacognitive monitoring while applying various cognitive skills. As shown in Table 3.5 and Table 3.6 (number of samples = 9859), distinct correlations were observed between the expression of facial cues and MMP performance.

For Task 1 (IC test), an increase in MMP values correlated with heightened average expression levels of cheek raising (AU_6) and blinking (AU_{45}). Specifically, it happens with cheek raising exhibits greater variability in expression, signifying more pronounced movements and more frequent blinking. Additionally, the direction of gaze predominantly shifted leftward with minimal deviation, reflecting a more stable and focused gaze to the left. The head pose was typically elevated and simultaneously tilted downward, suggesting a specific attentive stance adopted during this task.

For Task 2 (VSWM test), as MMP values escalated, there was a noticeable increase in the average expression levels of brow lowering (AU_4) and lips parting (AU_{25}). Brow lowering showed considerable variability, suggesting fluctuating intensity in frowning. Conversely, lip corner pulling (AU_{12}) demonstrated a decrease in average expression levels, implying less frequent smiling or positive expressions. This reduction was accompanied by lower variability in its expression, indicating a more consistent facial demeanor. Additionally, the movement of the gaze vertically and the head pose both horizontally and vertically showed reduced variability, indicating a more controlled and steady positioning throughout the task.

For Task 3 (TS test), an increase in MMP was associated with a heightened average expression level of lips parting (AU_{25}), indicating more frequent occurrences of this behavior. Conversely, lip corner pulling (AU_{12}) showed a decrease in average expression levels, accompanied by lower variability, suggesting less frequent smiling or positive facial cues. The lip corner depressor (AU_{15}) and blinking (AU_{45}) remained stable in average expression but exhibited contrasting variability patterns: the lip corner depressor showed lower variability, indicating a more consistent expres-

sion, while blinking displayed higher variability, reflecting more frequent blinking. In terms of gaze and head movement, there was lower variability in the change of the left-right gaze direction, suggesting a steadier horizontal gaze. The horizontal head pose changes showed higher variability, indicating more frequent adjustments, while changes in left-right head rotation exhibited lower variability, reflecting a more stable head orientation.

For Task 4 (IC & VSWM), the lip corner puller (AU_{12}) was the only facial cue to show a significant correlation with MMP. As MMP values increased, there was a noticeable decreasing trend in the average expression levels of AU_{12}, accompanied by lower variability. This pattern indicates a reduction in both the frequency and intensity of smiling or positive expressions, suggesting that increased metacognitive activity during the test is associated with decreased display of positive emotions.

For Task 5 (TS & VSWM), an increase in MMP correlated with a rise in the average expression level of lip parting (AU_{25}), indicating more frequent lip parting. In contrast, the average expression level of lip corner pulling (AU_{12}) decreased, accompanied by lower variability, suggesting a reduction in smiling or positive facial cues. Expressions such as outer brow raising (AU_2), dimpling (AU_{14}), and lip tightening (AU_{23}) showed no significant change in average levels but exhibited significantly lower variability, reflecting more consistent and stable expressions. Additionally, gaze direction and head pose were notably steadier, with head location moving away from the screen, and gaze direction remaining more stable throughout the task.

For Task 6 (IC & TS & VSWM), as MMP values increased, there was an enhanced stability in the expression of eyelid tightening (AU_7) and the horizontal gaze direction. This pattern suggests that with higher muscle potential, there was a significant reduction in the variability of eyelid tightening and a more consistent horizontal gaze.

The analysis above is used to reduce the scope of analyzing facial cues. Now, the results derive a set of facial cues that have a significant correlation to MMP. In Section 3.10, a regression analysis of facial cues is conducted to predict MMP in all tasks.

Table 3.5: Pearson-coefficients between facial cues' average level (ave.) and MMP. Note: * and yellow highlighting indicate statistical significance ($p < .05$). IC: inhibitory control; VSWM: visual and spatial working memory; TS: target switching.

(a) AUs

Task ID	**Cognitive tests**	**Ave. of AU_1 to AU_{12}**								
		AU_1	AU_2	AU_4	AU_5	AU_6	AU_7	AU_9	AU_{10}	AU_{12}
Task 1	IC	-0.12	-0.01	0.01	0.08	0.20*	0.09	0.12	0.12	0.02
Task 2	VSWM	0.00	0.06	0.21*	0.10	-0.06	0.01	0.02	-0.03	-0.17*
Task 3	TS	-0.12	0.04	0.06	0.05	-0.02	0.07	0.11	-0.01	-0.17*
Task 4	IC&VSWM	0.07	0.03	0.09	0.03	-0.08	-0.03	-0.02	-0.13	-0.20*
Task 5	TS&VSWM	-0.09	-0.08	0.07	0.09	-0.03	-0.01	0.08	-0.06	-0.15*
Task 6	IC&TS&VSWM	-0.10	0.08	0.15	0.08	-0.06	0.01	-0.04	-0.09	-0.08
		Ave. of AU_{14} to AU_{45}								
		AU_{14}	AU_{15}	AU_{17}	AU_{20}	AU_{23}	AU_{25}	AU_{26}	AU_{45}	
Task 1	IC	0.04	-0.03	-0.07	0.05	0.09	-0.13	0.04	0.19*	
Task 2	VSWM	-0.11	0.00	-0.10	-0.02	-0.03	0.15*	0.00	-0.02	
Task 3	TS	-0.06	-0.05	-0.06	-0.07	-0.02	0.19*	0.06	-0.02	
Task 4	IC&VSWM	-0.12	-0.01	-0.00	0.03	-0.00	0.01	0.04	-0.10	
Task 5	TS&VSWM	-0.14	0.08	-0.05	0.01	-0.11	0.17*	0.04	-0.05	
Task 6	IC&TS&VSWM	-0.13	-0.06	-0.06	-0.14	-0.01	0.08	0.08	-0.00	

(b) Gaze direction

Task ID	**Cognitive tests**	**Ave. of gaze direction (G)**	
		G_x	G_y
Task 1	IC	0.29*	-0.04
Task 2	VSWM	-0.04	-0.01
Task 3	TS	0.04	-0.13
Task 4	IC&VSWM	0.06	-0.02
Task 5	TS&VSWM	0.02	-0.01
Task 6	IC&TS&VSWM	0.03	0.05

Table 3.5: Pearson-coefficients between facial cues' average level (ave.) and MMP. Note: * and yellow highlighting indicate statistical significance ($p < .05$). IC: inhibitory control; VSWM: visual and spatial working memory; TS: target switching. (continued)

(c) Head pose

Task ID	**Cognitive tests**	**Ave. of head pose location (H(loc.)) and rotation (H(rot.))**					
		$H(loc.)_x$	$H(loc.)_y$	$H(loc.)_z$	$H(rot.)_x$	$H(rot.)_y$	$H(rot.)_z$
Task 1	IC	-0.09	-0.17*	-0.04	-0.16*	-0.07	-0.02
Task 2	VSWM	-0.08	-0.06	0.14	0.02	0.03	-0.07
Task 3	TS	0.01	-0.12	-0.09	-0.01	-0.05	0.06
Task 4	IC&VSWM	-0.04	-0.01	-0.04	0.15	-0.06	0.12
Task 5	TS&VSWM	-0.13	-0.08	-0.00	0.22*	-0.11	0.03
Task 6	IC&TS&VSWM	-0.04	0.09	0.01	0.15	-0.03	0.01

Table 3.6: Pearson-coefficients between facial cues' standard deviation (std) and MMP. Note: * and yellow highlighting indicate statistical significance ($p < .05$). IC: inhibitory control; VSWM: visual and spatial working memory; TS: target switching.

(a) AUs

Task ID	**Cognitive tests**	**Std. of AU_1 to AU_{12}**								
		AU_1	AU_2	AU_4	AU_5	AU_6	AU_7	AU_9	AU_{10}	AU_{12}
Task 1	IC	-0.06	0.01	-0.06	0.06	0.16*	-0.00	0.09	0.09	-0.06
Task 2	VSWM	-0.06	-0.00	0.17*	0.07	-0.06	-0.13	-0.07	-0.09	-0.22*
Task 3	TS	-0.11	0.01	-0.02	-0.03	0.02	0.05	0.02	-0.07	-0.17*
Task 4	IC&VSWM	-0.00	0.03	-0.04	0.03	-0.09	-0.09	-0.02	-0.15	-0.19*
Task 5	TS&VSWM	-0.12	-0.16*	-0.06	-0.00	-0.05	-0.14	-0.06	-0.15	-0.23*
Task 6	IC&TS&VSWM	-0.06	0.03	0.07	0.04	-0.02	-0.20*	-0.00	-0.13	-0.09
		Std. of AU_{14} to AU_{45}								
		AU_{14}	AU_{15}	AU_{17}	AU_{20}	AU_{23}	AU_{25}	AU_{26}	AU_{45}	
Task 1	IC	-0.05	-0.09	0.05	0.04	0.07	-0.14	0.01	0.12	
Task 2	VSWM	-0.10	-0.15	-0.08	-0.14	-0.11	-0.10	-0.11	-0.10	
Task 3	TS	-0.10	-0.19*	-0.05	-0.11	-0.00	-0.14	-0.10	0.17*	
Task 4	IC&VSWM	-0.13	-0.10	0.04	-0.01	-0.02	-0.12	-0.08	-0.13	
Task 5	TS&VSWM	-0.18*	-0.01	0.04	-0.12	-0.16*	0.02	-0.08	-0.08	
Task 6	IC&TS&VSWM	-0.11	-0.13	0.09	-0.04	0.02	-0.05	-0.06	-0.08	

(b) Gaze direction

Task ID	**Cognitive tests**	**Std. of gaze direction (G)**	
		G_x	G_y
Task 1	IC	-0.29*	-0.03
Task 2	VSWM	-0.11	-0.18*
Task 3	TS	-0.25*	0.09
Task 4	IC&VSWM	-0.11	-0.04
Task 5	TS&VSWM	-0.23*	0.01
Task 6	IC&TS&VSWM	-0.17*	-0.10

Table 3.6: Pearson-coefficients between facial cues' standard deviation (std) and MMP. Note: * and yellow highlighting indicate statistical significance ($p < .05$). IC: inhibitory control; VSWM: visual and spatial working memory; TS: target switching. (continued)

(c) Head pose

Task ID	Cognitive tests	Std. of head pose location (H(loc.)) and rotation (H(rot.))					
		$H(loc.)_x$	$H(loc.)_y$	$H(loc.)_z$	$H(rot.)_x$	$H(rot.)_y$	$H(rot.)_z$
Task 1	IC	-0.03	-0.01	0.02	0.07	-0.13	-0.07
Task 2	VSWM	-0.04	-0.16*	-0.17*	-0.14	-0.14	-0.15
Task 3	TS	0.17*	0.15	0.06	0.11	-0.16*	0.10
Task 4	IC&VSWM	-0.02	0.05	-0.06	0.08	-0.04	0.04
Task 5	TS&VSWM	-0.09	0.02	-0.02	0.00	-0.14	-0.10
Task 6	IC&TS&VSWM	0.04	-0.08	0.02	-0.00	-0.05	0.04

3.9 Discussion

Building on the results presented in Section 3.8, this discussion focuses on two key areas. First, it examines the reinforcement of the positive relationship between MMP and cognitive task performance in young learners aged 7 to 11, specifically within mathematics-related tasks conducted in CBLEs. Second, it explores the potential of predicting MMP through pupils' facial cues, offering new insights into non-intrusive approaches for metacognitive monitoring.

3.9.1 MMP Predicts Task Performance in Pupils

Overall, the data consistently support the hypothesis that higher metacognitive monitoring capabilities (TCI) are associated with improved task performance (TS). Across all tasks, the average increase in task scores is approximately 4.09 points for each one-unit increase in TCI.

This study reinforces prior research demonstrating a strong positive correlation between MMP and task performance, confirming that young pupils with higher MMP consistently achieve better learning outcomes (Higgins et al. 2016, Bellon et al. 2020,

Kautzmann & Jaques 2019). Importantly, this study extends these findings to scenarios where young pupils engage in mathematics-related cognitive tasks within CBLEs. These results offer insights for future research aiming to enhance pupils' learning outcomes in CBLEs by improving their metacognitive monitoring abilities.

3.9.2 Impact of Pupils' MMP on Facial Cues

While prior studies have demonstrated the correlation between facial cues and MMP in undergraduate populations (Taub et al. 2018, 2021), this study is the first to establish this correlation in younger learners, specifically pupils aged 7 to 11 in a CBLE.

The findings reveal that multiple facial expressions and head movements are significantly associated with MMP in this age group, extending previous research from undergraduate students to an early education group. Furthermore, this study reveals that the strength of these correlations is influenced by the gender factor, providing new insights into how individuals shape the relationship between facial cues and MMP.

In particular, increases in MMP were consistently associated with heightened expressions in certain muscle groups-most notably increased lip parting-coupled with a reduction in the frequency and variability of expressions typically linked to positive affect, such as those involving the lip corner puller. This shift away from overtly positive expressions may indicate that effective MMP is less about maintaining a pleasant effect and more about engaging cognitive resources that could include subtle signals of frustration or focused concentration.

These facial adjustments appear to be part of a broader pattern wherein more stable and controlled gaze directions and head movements are observed alongside changes in facial muscle activity. Such steadiness likely reflects an underlying focus of attention that is critical for accurately monitoring one's cognitive processes. The reduction in variability of certain expressions (for example, a more consistent eyelid tension and less frequent smiling) points to a state of engagement that minimizes distractions like boredom Cloude et al. (2020).

Furthermore, the results align with the findings in undergraduate students Taub et al. (2021), who demonstrated that while unexpected affective responses (e.g., surprise) tend to disrupt metacognitive accuracy, a controlled level of frustration may enhance monitoring performance. The observed decrease in positive, smiling expressions, alongside increased indicators of cognitive effort (e.g., heightened lip

parting and controlled facial cues), suggests that young learners may adopt a more effortful, potentially slightly frustrated, stance that promotes more reflective and accurate self-assessment of their knowledge.

3.10 Preliminary Experiments for Estimating MMP in Following Chapter

3.10.1 Interaction Effects of Gender on Correlation between Facial Cues and MMP

Given that gender significantly influences facial expressions, with women typically exhibiting more pronounced positive expressions than men (Dimberg & Lundquist 1990), an analysis was conducted to re-examine the relationship between facial cues and MMP. This investigation specifically explored how this relationship differs between male and female participants and will contribute to next MMP estimations.

Significant interactions were observed between facial cues and MMP with respect to gender, highlighting distinct expression patterns associated with metacognitive processes in male and female pupils, see Table 3.7.

For male pupils, several facial cues exhibited significant positive correlations with MMP, including AU_4 (brow lowerer), AU_7 (lid tightener), AU_{23} (lip tightener), and AU_{25} (lips part). These results suggest that increased activation of these facial actions is associated with higher metacognitive monitoring performance among males. Conversely, some facial cues showed significant negative correlations, such as AU_{12} (lip corner puller), AU_{15} (lip corner depressor), AU_{45} (blink), and $H(loc.)_x$ (horizontal head location), indicating that these behaviors are related to lower MMP in male pupils.

For female pupils, the patterns were different. Only G_x (horizontal gaze direction) showed a significant positive correlation with MMP, suggesting that greater horizontal gaze shifts are linked to better metacognitive performance. Significant negative correlations were found for AU_{12} (lip corner puller), G_y (vertical gaze direction), and $H(loc.)_y$ (vertical head location), implying that frequent activation of these cues corresponds to lower MMP in females.

It is important to note that no facial cue showed an opposite correlation of di-

Table 3.7: Pearson-coefficients between average expressions of facial cues and MMP by gender. Note: * and yellow highlighting indicate statistical significance ($p < .05$).

Facial Cue	Males	Females
AU_4	0.06*	-0.02
AU_7	0.04*	-0.02
AU_{12}	-0.04*	-0.05*
AU_{15}	-0.04*	-0.02
AU_{23}	0.04*	-0.00
AU_{25}	0.06*	-0.02
AU_{45}	-0.04*	-0.02
G_x	-0.00	0.04*
G_y	-0.01	-0.05*
$H(loc.)_x$	-0.04*	0.00
$H(loc.)_y$	-0.01	-0.04*

rections between male and female pupils. However, gender appears to moderate the strength and significance of these correlations:

- Significant only for males: AU_4, AU_7, AU_{23}, AU_{25}, AU_{15}, AU_{45}, and $H(loc.)_x$.
- Significant only for females: G_x, G_y, and $H(loc.)_y$.
- Significant for both males and females: AU_{12} (both show a negative correlation).

Together, these findings highlight that gender is a key moderating factor in the relationship between facial cues and metacognitive monitoring. This insight is critical for developing more accurate, personalized models of MMP estimation, which will be addressed in the following chapter.

3.10.2 Multivariate analysis: How Multiple Facial Cues Together Predict MMP?

Building on prior studies that examined correlations between facial cues and MMP, primarily in adult learners, this section advances the research by investigating the

predictive power of facial cues for estimating MMP in pupils aged 7 to 11 within CBLEs. A multiple regression analysis is also conducted to identify which facial cues predict pupils' MMP. This section is the first to reveal both the performance of MMP estimation in young learners and the extent to which facial cues can explain and predict MMP.

OLS regression (Dismuke & Lindrooth 2006) was used to model MMP as a function of facial expression metrics. Prior to analysis, data were checked for normality, and outliers were addressed. The model included 50 predictors, encompassing both mean and standard deviation values of facial cues and head pose angles. The model achieved an R-squared of 0.02 and an adjusted R-squared of 0.02, indicating a modest fit[5]. The overall model was statistically significant ($p < 0.01$), suggesting that the set of predictors contributes to explaining variability in MMP.

Table 3.8: Regression analysis of facial cues that significantly predict MMP (TCI). Note: this table includes facial only cues with $p < 0.05$.

Facial Cues	**Coefficient**	**Standard Error**	**95% Confidence Interval**
Ave(AU_1)	-0.10	0.06	[-0.22, 0.02]
Ave(AU_4)	0.08	0.04	[0.01, 0.15]
Ave(AU_{25})	0.09	0.04	[0.00, 0.17]
Ave(head hori. pose)	-0.10	0.04	[-0.18, -0.03]
Ave(head vert. pose)	-0.08	0.04	[-0.16, -0.01]
Ave(Nodding)	0.10	0.05	[0.01, 0.19]
Ave(vert. gaze)	-0.10	0.05	[-0.21, 0.00]
Std(AU_6)	0.33	0.10	[0.14, 0.52]
Std(AU_{12})	-0.33	0.10	[-0.52, -0.14]
Std(hori. gaze)	-0.52	0.13	[-0.76, -0.27]
Std(vert. gaze)	0.35	0.14	[0.07, 0.62]

Among the facial cues analyzed, a select few had significant impacts on the MMP as illustrated in Table 3.8. Visualized facial AUs are shown in Figure 3.12. Notably, the average intensity of brow lowering (AU_4) demonstrated a positive correlation with

[5]R-squared measures the proportion of variance in the dependent variable explained by the independent variables in an OLS regression. It ranges from 0 to 1, where higher values indicate better explanatory power of the model.

MMP, where an increase in brow lowering intensity led to a rise in MMP scores, as evidenced by a coefficient of 0.08 ($p = 0.02$). Similarly, the intensity of lips parting (AU_{25}) was positively associated with MMP, increasing the score by 0.09 for each unit increase ($p = 0.05$). Head movements also played a crucial role, with both horizontal and vertical translations of the head showing negative associations with MMP, decreasing the score by 0.10 ($p = 0.01$) and 0.08 ($p = 0.03$), respectively, per unit increase. Additionally, the rotational motion around the x-axis, indicative of nodding, positively impacted MMP, increasing it by 0.10 with each unit increase ($p = 0.04$). The variability in expressions such as cheek raising (AU_6) and lip corner pulling (AU_{12}) also significantly affected MMP, with the standard deviation of AU_6 increasing MMP by 0.33 ($p < 0.01$) and that of AU_{12} decreasing it by 0.33 ($p < 0.01$). Lastly, gaze behavior was influential, where the standard deviation in the horizontal gaze angle significantly reduced MMP by 0.52 ($p < 0.01$), while the vertical gaze variation increased it by 0.35 ($p = 0.02$).

The most significant predictors of MMP appear to be gaze behaviors and head gestures, alongside specific facial action units. For example, greater variability in horizontal gaze direction has a strong negative effect on MMP, suggesting that frequent shifts in eye movement from side to side may indicate distraction or reduced attentional focus, which hinders metacognitive monitoring. In contrast, variability in cheek raising and lip corner pulling positively correlates with MMP. These variations may reflect dynamic emotional expressions, such as amusement, contentment, serenity, or even mild embarrassment, which have been linked to increased cognitive engagement and self-awareness (Ekman & Friesen 1978). This suggests that certain expressive facial movements may signal a more engaged or reflective mental state, based on which, learning technologies can be used to enhance MMP.

These findings illustrate the key facial expressions, head gestures, and gaze directions in predicting MMP, which were used for the subsequent classification attempt.

3.10.3 ML-based MMP Classification

Given the significant features identified in previous linear regression analyses, the performance of a machine learning model in classifying MMP into three classes was investigated. Three classes of MMP are: `precise` MMP ($ACI = 1$), `uncertain`

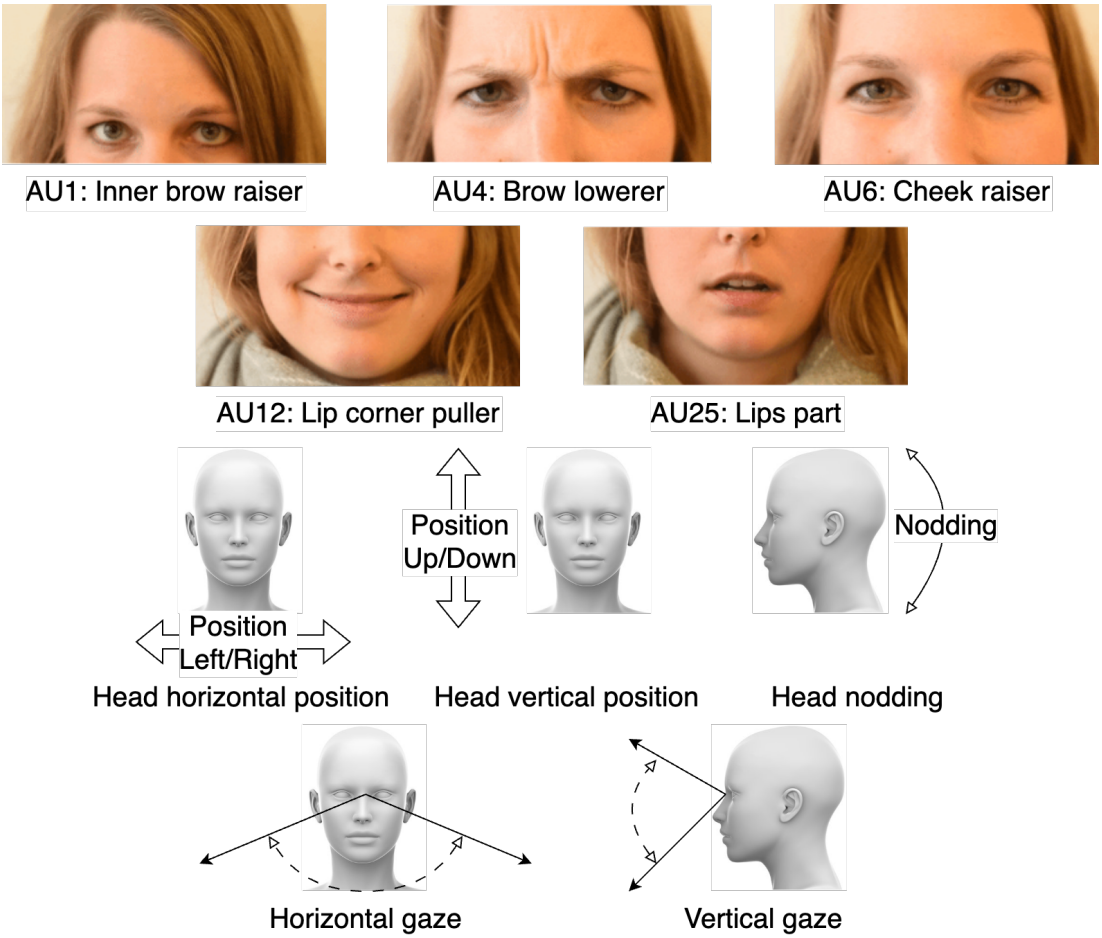


Figure 3.12: Key facial cues in OLS regression. Note: The AUs' animations are from iMotions (2018).

MMP ($ACI = 0.75$)[6], and `imprecise` MMP ($ACI = 0$). Given the multidimensional nature of the facial cues and behavioral metrics involved, selecting a robust classifier capable of handling high-dimensional spaces and complex decision boundaries is crucial. For this task, the Support Vector Machine (SVM) classifier (Hearst et al. 1998) was selected (the introduction of SVM was presented in Section 2.6.1). SVMs are particularly effective in classification tasks involving high-dimensional spaces, even when the number of dimensions exceeds the number of samples, making them well-suited for the feature-rich data in this study.

The SVM model was established to understand the separability of the classes based on the selected features in Table 3.8. A standard train-test split was employed, with 20% of the dataset reserved for testing to evaluate the model's performance. The data was standardized to ensure that all features contributed equally to the model, avoiding bias toward features with inherently larger scales. The trained SVM classifier achieved an accuracy of approximately 59.8% on the test set. This performance indicates a moderate level of capability in classifying the MMP levels using the identified facial cues. More advanced methods, such as deep learning, may better capture the nuanced patterns and improve classification performance. This will be further explored in the next chapter through the development of a neural network-based model designed to estimate MMP.

3.11 Chapter Summary

From experiments addressing RQ1.1, the results in the user study reinforced the importance of MMP in young pupils in CBLEs and explored the potential for identifying pupils' MMP through facial cues.

From experiments addressing RQ1.2, a set of facial expressions, including outer brow raiser, brow lowerer, cheek raiser, lid tightener, lip corner puller, dimpler, lip corner depressor, lip tightener, lip part, and blink, were found to be associated with pupils' MMP. Additionally, gaze direction, head position, and head movements (up/down, left/right) were also linked to variations in MMP. Notably, the inner brow raiser, brow lowerer, lip part, cheek raiser, lip corner puller, and both vertical and horizontal movements of the head and gaze were found to have predictive value for MMP.

These findings highlight the importance of pupils' MMP in CBLEs and the po-

[6] As the square operation in Equation 3.4, the ACI of 'Not sure' JOC is 0.75.

tential use of facial cues for estimating MMP. Facial cues can offer indirect insights into pupils' cognitive states, which is particularly valuable for developing adaptive and responsive CBLEs for enhancing MMP.

Furthermore, gender emerged as a key factor influencing this relationship, suggesting that facial cues linked to metacognitive states may differ between males and females. Especially, from the ML-based MMP classification study, its results indicate that relying solely on simple linear relationships is insufficient for accurately estimating MMP, and more advanced modeling approaches are needed to capture the complex, non-linear interactions between facial cues and metacognitive monitoring processes.

As a baseline to support the long-term aim of improving children's mathematics learning outcomes through enhanced metacognitive monitoring (as detailed in Chapter 5), the work of this chapter paves the way for future automatic metacognitive monitoring supports in educational technologies for young children's mathematics learning (Harter 2012, Harris & Brown 2013, Lehnert 2024).

To be clear, this work makes four key contributions:

1. Reinforces the positive impact of MMP on young pupils' mathematics-related task performance in CBLEs, providing empirical evidence from children aged 7 to 11.

2. Identifies specific facial cues that are indicative of MMP in children, advancing the understanding of behavioral markers relevant for metacognitive state estimation in this age group.

3. Shows that simple linear models are insufficient for accurately capturing the complex relationship between facial cues and MMP, highlighting the need for more sophisticated modeling approaches.

4. Reveals that gender significantly moderates the correlation between MMP and facial cues, emphasizing the importance of considering demographic factors when developing predictive models of metacognitive monitoring.

Chapter 4

Nurturing Self-aware Learning through Facial Expression Interpretation

---

纸上得来终觉浅，绝知此事要躬行。(Translation: Knowledge gained from paper always feels superficial; To truly grasp a matter, one must personally practice it.)

— 陆游 (Lu You)

This chapter reports an approach to estimate metacognitive monitoring performance (MMP) through facial expressions. Content from this chapter has been presented at the ACM CHI 2025 in the paper Nurturing Self-aware Learning through Facial Expression Interpretation (Ruan, X., Constantin, A., Palansuriya, C., Wang, K. & Atkinson, M., 2025, April).

Learning technologies are designed to tailor support to improve learning by recognizing and responding to learners' feelings about learning tasks. An advanced technique, Meta-Facial Expression Interpreter (M-FEI), is proposed in this chapter's work for estimating `precise`, `imprecise`, and `uncertain` MMP in pupils, which were illustrated in Section 3.10.3.

In addition to the user study of Chapter 3 in China, an additional user study was conducted in Scotland to validate the generalizability of M-FEI. Through M-FEI, pupils' MMP was estimated by analyzing their facial expressions. This approach represents a significant improvement over a conventional method, offering a more immediate response when a learner encounters a new task.

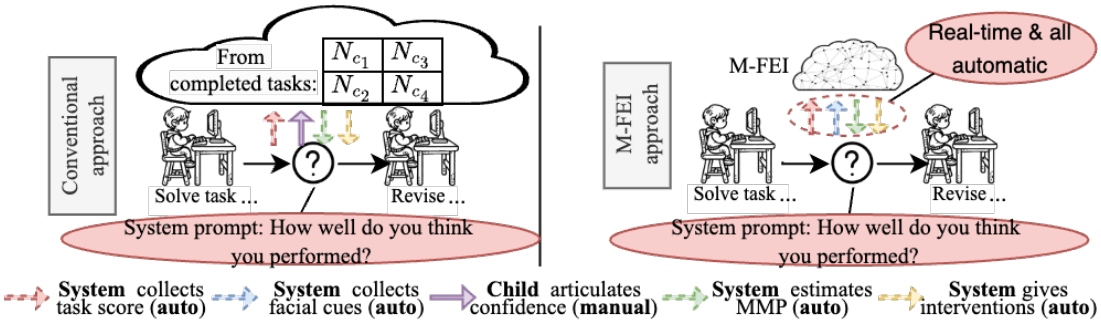


Figure 4.1: Comparison of interventions using two approaches of estimating MMP (conventional approach and the Meta-Facial Expression Interpreter (M-FEI)). The green arrow of the conventional approach is the estimated MMP based on numbers N_{c_1} to N_{c_4}. Specifically, N_{c_1} is the number of instances that the pupil judged their answer as correct and solved the task correctly; N_{c_2} is the number of instances that the pupil judged their answer as correct but solved the task incorrectly. Similarly, N_{c_3} counts instances when pupils thought they were wrong yet were correct, and N_{c_4} counts cases where pupils thought they were wrong and they were. The green arrow of M-FEI represents the MMP estimated based on facial expressions while pupils are reflecting on their solutions. The details are presented in Sections 4.6.1 and 4.6.2.

4.1 Introduction

As introduced in Chapter 2, there has been a significant shift towards computer-based learning environments (CBLEs) in education in recent years. Advanced technologies demonstrate that tailoring metacognitive interventions using the conventional approach, like the knowledge monitoring assessment (KMA) matrix, improves learning outcomes (Kautzmann et al. 2016, Kautzmann & Jaques 2019, Guo 2020).

However, previous research has raised concerns about the accuracy of conventional approaches and their difficulties in articulating confidence for young learners (Pintrich et al. 2000, Mihalca & Mengelkamp 2020). Stemming from the link between facial cues and MMP, this chapter introduces a deep-learning approach called Meta-Facial Expression Interpreter (M-FEI)[1] for enhancing MMP estimation (Taub & Azevedo 2018, Taub et al. 2021). This work has developed M-FEI for tailoring metacognitive

[1]In the context of M-FEI, 'facial expressions' refers to the visual appearance of the face in recordings, rather than the underlying facial muscle movements typically referenced in the definition chapter. The term 'expression' is used here because, in the machine learning field, it is commonly adopted to describe facial appearances in analysis.

monitoring support. An example comparing M-FEI with the conventional approach is presented in Figure 4.1. For instance, when M-FEI identifies imprecise metacognitive monitoring in a learner, it triggers the targeted strategy to provide tailored hints and examples, aiding learners in assessing their knowledge (Kautzmann et al. 2016, Kautzmann & Jaques 2019). Conversely, when learners display confidence precisely, the system minimizes guidance, fostering autonomy and maintaining the learning flow. This automatic estimation of MMP maintains learners' flow and engagement by eliminating the need for confidence articulation (Amershi et al. 2014, Komatani & Nakano 2020) and enables real-time tailoring of metacognitive monitoring interventions for individual needs.

A user study was conducted in two countries to develop a dataset of pupils' facial expressions, named Affect2Metacognition (A2M), while they engaged in metacognitive monitoring. The M-FEI was developed to estimate pupils' MMP by interpreting facial expressions, and this method was compared with the conventional approach. The comparison revealed limitations in the accuracy of the conventional approach for MMP estimation. In contrast, the M-FEI method achieved an average 14% increase in accuracy for identifying MMP, along with a 15% reduction in false alarms.

4.2 Research Questions

Following RQ1.1 and RQ1.2 from Chapter 3, the current chapter investigates the identification of MMP and addresses the following RQ1.3 and RQ1.4.

4.3 Related Work Informing M-FEI Design

This chapter presents a technique, M-FEI, designed to estimate learners' MMP and tailor metacognitive monitoring interventions. Accordingly, the related work section will review three key areas: research on adaptive support for metacognitive monitoring, the benefits of eliminating self-articulation in MMP estimation within CBLEs, and the feasibility of applying deep learning techniques in M-FEI.

4.3.1 Adaptive Support of Metacognitive Monitoring

Considering the significant impact of MMP on pupils' learning outcomes, extensive research has focused on developing interventions that enhance MMP during the completion of educational tasks. Barzilai et al. (Barzilai & Blau 2014) showed that metacognitive instructions, which introduce learning concepts and variables, have a positive impact on children's performance. Prior research includes investigations into the effectiveness of metacognitive support, using feedback on achieved performance, for pupils' metacognitive monitoring. Works from Urban & Urban (2021*b*,*a*) highlight how feedback on past performance can positively influence the student's MMP. Prompting pupils with instructions about their past performance has been shown to significantly boost their MMP (Urban & Urban 2021*b*).

Despite the benefits of providing performance feedback, there is evidence suggesting that such instructions may lead pupils to rely excessively on the feedback when reflecting on their performance, potentially at the expense of neglecting other relevant task information, such as a task's specific details and objectives (Epley & Gilovich 2006, Simmons et al. 2010, Urban & Urban 2021*a*). This reliance could limit the development of metacognitive skills. To address these challenges, studies have focused on developing adaptive support for metacognitive processes, which are gradually withdrawn/increased based on the estimations for occurrences of their `imprecise` MMP. For example, Kautzmann et al. (Kautzmann et al. 2016, Kautzmann & Jaques 2019) introduced an Animated Pedagogical Agent (APA), as previously discussed in Section 2.4.2. APA agent prompts pupils to articulate their confidence and uses that input to adapt subsequent metacognitive monitoring support. Similarly, research by Guo et al. (Guo 2020) also employs a variant of the KMA to estimate MMP, relying on historical MMP data to prompt different information while subjects are engaged in learning tasks. These studies claimed that better learning performance was achieved by learners with real-time adaptive support, which underscores the benefit of real-time adaptation of metacognitive support based on the conventional MMP estimation (Kautzmann et al. 2016, Kautzmann & Jaques 2019, Guo 2020).

4.3.2 Conventional MMP Estimation and Automatic MMP Estimation

As observed in the previous section, conventional MMP estimation typically relies on results from KMA based on previously completed tasks. However, this approach has substantial limitations. Firstly, the relevance of past KMA outcomes in estimating MMP for upcoming tasks diminishes when the new tasks significantly differ in content or domain. For instance, a pupil's previous performance in reading comprehension tasks may not accurately estimate their MMP in arithmetic tasks (Pintrich et al. 2000, Mihalca & Mengelkamp 2020). Secondly, variations in a pupil's prior knowledge and experiences can skew MMP classifications, particularly when shifting from one topic to another (Mihalca & Mengelkamp 2020). Thirdly, reliance on self-assessment poses challenges, particularly for younger learners. For instance, younger elementary school students tend to provide less nuanced self-descriptions and often lack the metacognitive skills necessary to accurately identify areas of knowledge, especially when tackling complex tasks (Harter 2012, Harris & Brown 2013, Amershi et al. 2014, Lehnert 2024). Furthermore, repeated self-assessment inquiries can disengage learners from the learning content (Komatani & Nakano 2020).

Recent advances in HCI advocate automatic methods that respond to spontaneous human behaviors, such as eye movements, facial expressions, and interaction patterns, to monitor learners' processes in real-time (D'Mello et al. 2007, Baker & Ocumpaugh 2014). Vanneste et al. (Vanneste et al. 2021) implemented a computer vision-based model in classrooms that can recognize behaviors indicative of cognitive engagement, such as note-taking or hand-raising, without requiring learners to verbally describe their activities. Similarly, Behera et al. (Behera et al. 2020) explored the automatic detection of non-verbal behaviors, including hand-over-face gestures, and head and eye movements, as well as emotions expressed through facial expressions. Their method effectively identified when learners were struggling with difficult tasks, bypassing the need for them to articulate their difficulties explicitly. Furthermore, Baltaci et al. (Baltaci & Gokcay 2016) monitored students' stress levels by assessing pupil dilation and facial temperature, showing the potential of interpreting learners' psychological signals. These findings underscore the feasibility of an automatic approach to MMP estimation that relies on real-time, spontaneous behaviors.

Prior work has shown that explicitly alerting users is not always an optimal inter-

vention. The proposed M-FEI, to estimate MMP using facial expressions, represents a non-intrusive approach termed 'mindless' (Riku 2021) that improves the flow of learning. In this way, from the perspective of user experience in computer-based learning environments (CBLEs), M-FEI not only provides automatic MMP estimation, but also enables learners to focus on the flow of learning by removing the need to articulate their feelings (Kautzmann et al. 2016, Kautzmann & Jaques 2019, Guo 2020). Consequently, the automatic nature method is crucial as it avoids distractions and enables learners to concentrate on learning.

4.3.3 Relationship between Facial Cues and Metacognitive Monitoring

This thesis reviewed prior works that investigate links between learners' feelings and facial expressions, and their MMP, in Chapter 2. Thus, in this review section, the details of those works will not be reviewed.

4.3.4 Potential of Deep-learning in MMP Estimation

Recent work in interpreting facial expressions to assess learners' self-reflection is based on the assumption that a learning environment that is sensitive to learners' affective states enriches learning (D'Mello et al. 2007, 2009, Lepper & Henderlong 2000, Graesser et al. 2007). Research indicates that novice human tutors are unable to accurately assess a student's understanding based on their facial expressions (Person et al. 1994, Chi et al. 2004, D' Mello & Graesser 2012). Instead, they tend to make approximations rather than precise and detailed assessments. In contrast, experienced teachers are better able to discern students' understanding through facial expressions, effectively using this information to assess individual needs, identifying those requiring additional support, and adjusting the pace or content of the learning material accordingly (Behera et al. 2020). However, computers could potentially outperform, particularly novice human tutors, where artificial intelligence algorithms can accurately monitor learners' behaviors (D'Mello & Graesser 2013). Recent research has assessed approaches to interpreting facial expressions during learning. For example, Kort et al. (Kort et al. 2001) proposed a comprehensive four-quadrant model that was used in their affective learning companion. This model recognizes facial expressions by monitoring facial features. D'Mello et al. (D' Mello & Graesser 2012) trained a

dynamic decision network to interpret facial data to infer students' cognitive states during learning.

Deep learning has significantly advanced facial expression recognition, particularly for emotions like confusion, frustration, and boredom, which are common during learning processes. Such technologies are increasingly crucial in educational settings, and Singh et al. (Singh et al. 2023) used deep-learning networks (Arnab et al. 2021), achieving a 66.5% accuracy rate, which is a 4.66% improvement over previous models. Harley et al. (Harley et al. 2015) found a 60.1% correlation between facial expressions detected by FaceReader (*FaceReader (5.0)* 2025) and physiological responses measured by electrodermal activity. Meanwhile, Aly (Aly 2024) demonstrated the effectiveness of pre-trained deep-learning networks when identifying learning-related emotional states, providing reliable affect monitoring in educational environments. These findings highlight deep learning's potential to enhance educational technologies by accurately interpreting learner emotions.

In summary, the relationship between affective states and metacognitive monitoring, combined with the feasibility of interpreting facial expressions using deep neural network-based AI, highlights the value of further investigating deep learning approaches for identifying MMP and tailoring metacognitive monitoring supports in CBLEs.

4.4 Additional Data Collection User Study in Scotland

As described in Chapter 3, the initial phase of the data collection was conducted at two primary schools in the eastern (Region A) and middle (Region B) parts of China. To assess M-FEI's generalizability for future researchers, another phase of this research was conducted in Scotland (Region C), UK, to collect data distinct from that in China. These regions were deliberately selected to capture notable contrasts, particularly between the UK and China, in terms of culture, traditions, economic status, degree of urbanization, and even common facial expressions (Samizadeh 2022, Yonglan et al. 2023).

Ethical approval for the additional user study in Scotland was granted under the same application as the user study described in Chapter 3, approved by the

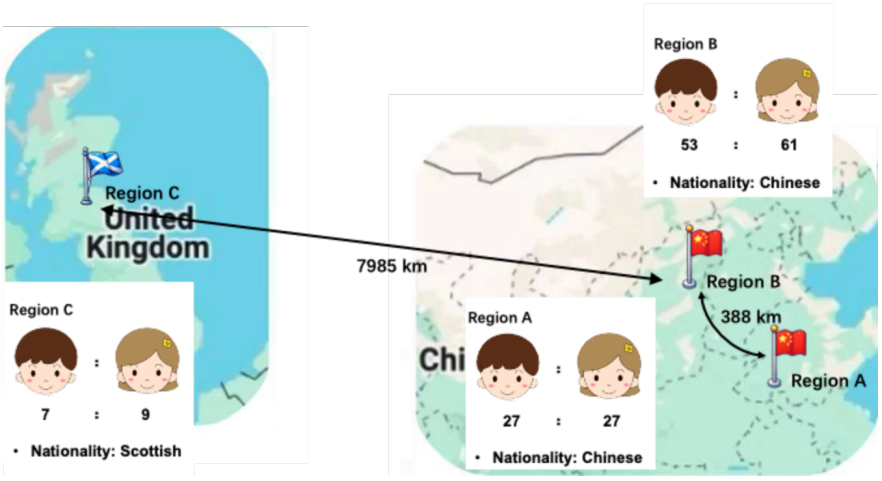


Figure 4.2: Location map of the first user study highlighting the second phase.

Ethics Approval Department of the University of Edinburgh, Department of Informatics (RT#6963) in November 2023. Participation in the extra user study was strictly voluntary, where Scottish pupils, together with their guardians, signed both pupils' and parents' participant information sheets and the informed consent forms[2]. Before enrollment, a comprehensive introduction to the user study was provided, and relevant documents were distributed. Participants were assured that their involvement would not affect their grades or activities at school. Additionally, the data management policy and participants' rights within the user study were clearly explained, with particular emphasis on their right to discontinue participation or withdraw their data at any time without any consequences. All collected data, including videos and generated logs, are securely stored on a password-protected, encrypted server and removed within one week (outlined in the data management policy). Access to the data was restricted to the principal researcher (PR), the author of this dissertation. To support future research, statistical measurements of facial cues and generated logs have been shared on GitHub (github/affect2mmp 2025), enabling researchers to verify the results or build upon the analyses.

A summary of the participants in the first user study is presented in Table 4.1. The first phase of the user study in China involved data collection from 168 pupils, as described in Chapter 3. For the second phase in Scotland, 16 pupils (7 males, 9 females; mean age = 9.6, SD = 1.8) participated in the user study[3]. Therefore, the

[2]A copy version of these documents is available on the GitHub site, (github/affect2mmp 2025).

[3]The number of pupils from Scotland is relatively small compared to the China dataset because the primary purpose of the user study in Scotland was to validate the M-FEI model rather than to

Table 4.1: Participant distribution across data collection phases and regions

Phase	Region	Total Pupils	Gender (Male / Female)
First Phase (China)	Region A	114	53 / 61
First Phase (China)	Region B	54	27 / 27
Second Phase (Scotland)	Region C	16	7 / 9

combined dataset for the first user study consisted of data from 184 pupils (87 males, 97 females; mean age = 9.81, SD = 1.37) from China and Scotland, see Table 4.1. The location map of those three regions is marked in Figure 4.2.

4.5 Affect2Metacognition Dataset

This section describes the development of the A2M dataset. It starts with the data annotation for the five classes of the MMP. Then it describes the pre-processing procedure for increasing the quality of the data. Given the potential for data bias on a deep-learning neural network, the last part of this section reports the demographic compositions and demographic bias in A2M.

4.5.1 Data Annotation

Overall, two types of data were extracted: (i) videos of facial expressions; and (ii) log files of the Meta-Brainhood software application that was described in the previous chapter. Timestamps from log files were used to index the corresponding video clips during which pupils completed their JOC questions. After this process, video clips of pupils' MMP were derived.

The attempt score (AS) and the attempt confidence rating (ACR) of attempt p in task t have been described, see Section 3.6.1.

For clarity, this chapter lists all cases of MMP. The MMP for attempt p in task t was calculated using $ACR_{t,p}$ and $AS_{t,p}$. The MMP has five classes, as illustrated in Figure 4.3. Specifically, the '+ +' class indicates that the pupil judged the answer as correct ($ACR_{t,p} = 1$), and the answer was indeed correct ($AS_{t,p} = 1$). The '+ -' class

collect training data. Additionally, due to time constraints in participant recruitment, the user study ultimately involved 16 pupils — the same number as in the Chinese test dataset for fair comparison.

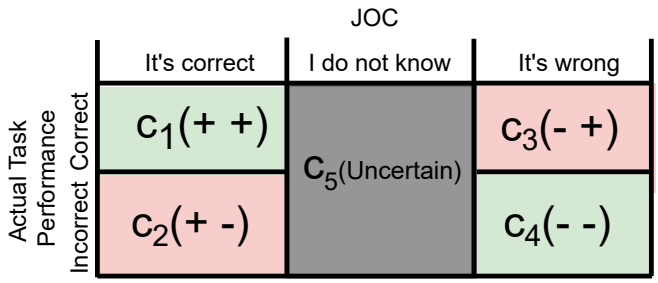


Figure 4.3: There are five classes of MMP: c_1 and c_4 represent `precise` MMP (green), c_5 represents `uncertain` (grey), c_2 and c_3 represent `imprecise` MMP (red).

applies when the pupil judged the answer as correct ($ACR_{t,p} = 1$), but the answer was incorrect ($AS_{t,p} = 0$). Conversely, the '- +' class refers to situations where the pupil judged the answer as wrong ($ACR_{t,p} = 0$), yet the answer was correct ($AS_{t,p} = 1$). The '- -' class is used when the pupil judged the answer as wrong ($ACR_{t,p} = 0$), and the attempt was also solved incorrectly ($AS_{t,p} = 0$). Finally, the 'uncertain' class is used when the pupil selected 'I do not know' ($AS_{t,p} = 0.5$). Following this, a video clip was categorized as a `precise` MMP if the pupil's confidence aligned with the attempted score (either c_1 or c_4); it was categorized as an `imprecise` MMP (either c_2 or c_3), as in (Kautzmann et al. 2016, Kautzmann & Jaques 2019); additionally, a video clip was categorized as an `uncertain` MMP if the pupil could not articulate their confidence (c_5).

4.5.2 Dataset Filtering and Dataset Preparation

In the data collection with Meta-Brainhood, some pupils appeared to become bored and disengaged, repeatedly and rapidly selecting the same JOC option. To improve the quality of the data for training purposes, a time filter was implemented to exclude such short video clips (2,732 clips). This refinement reduced the A2M dataset to 8,308 video clips, with clip durations ranging from 1.50 to 9.95 seconds.

The A2M dataset was labeled based on existing strategies to enhance metacognitive monitoring (Kautzmann et al. 2016, Kautzmann & Jaques 2019), classifying MMP into `imprecise` (c_2 or c_3), `precise` (c_1 or c_4), and `uncertain` (c_5). For the training and validation phases, clips from 152 pupils in China were randomly selected.

Table 4.2: The number of clips in the A2M dataset. For a demographic comparison of these datasets, see Figure 4.4. Note, % means the percentage of clips in the corresponding dataset.

Training (China)			**Validation (China)**		
Precise	Imprecise	Uncertain	Precise	Imprecise	Uncertain
3543 (57%)	2084 (33%)	617 (10%)	386 (56%)	244 (35%)	64 (9%)
Test (China)			**Test (Scotland)**		
Precise	Imprecise	Uncertain	Precise	Imprecise	Uncertain
440 (64%)	207 (30%)	40 (6%)	491 (72%)	124 (18%)	68 (10%)

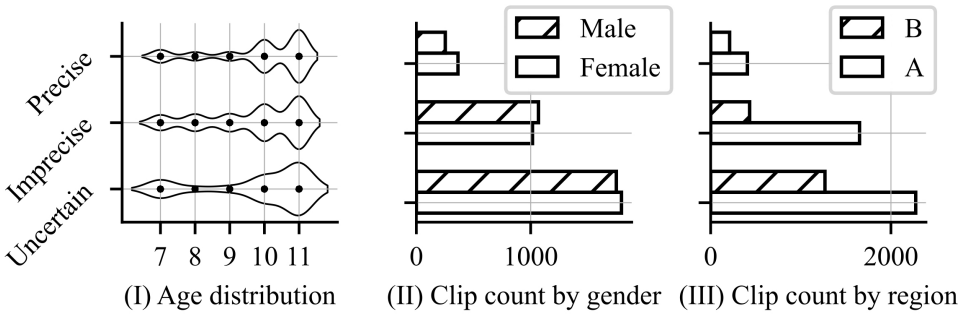
Their video clips were assigned to the training dataset (90%) and the validation dataset (10%). The remaining 16 pupils from China and 16 pupils from Scotland formed two distinct test datasets[4]. This structured partitioning ensures effective model training, validation, and testing by providing distinct datasets that minimize similarities, a practice that is well-supported in the literature on training neural networks (Li & Deng 2020). This data split is crucial for assessing the model's real-world applicability, reflecting the proposed approach's generalization of MMP estimation in practical scenarios. The distribution of clips across MMP categories is visualized in Table 4.2.

4.5.3 Demographic Bias of A2M and Data Curation

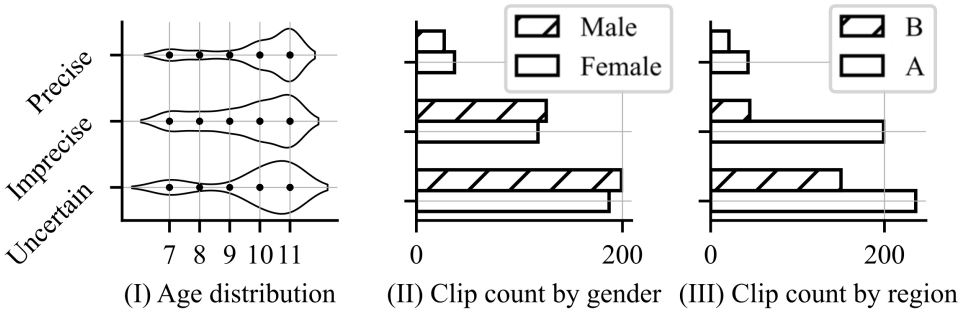
Deep learning is a data-driven approach; the composition of its training dataset often determines its outcomes. Recently, there has been an increased focus on data bias issues, particularly concerning the social dynamics of power embedded in data (Leavy et al. 2021). In this section, the evaluation examined the A2M's fairness conditions based on the recent research on ethical data curation for AI (Leavy et al. 2021, Kleinberg et al. 2016).

The demographic distribution in terms of age, gender, and geographical region within the A2M dataset was examined, as shown in Figure 4.4. The dataset includes

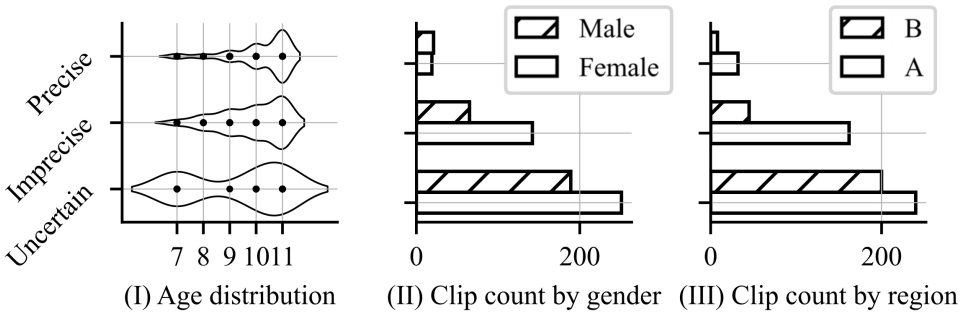
[4]In compliance with the Scottish ethical approval and as stated in the participant information sheet, pupil data from Scotland is restricted solely for validation purposes and was not used for training deep learning models.



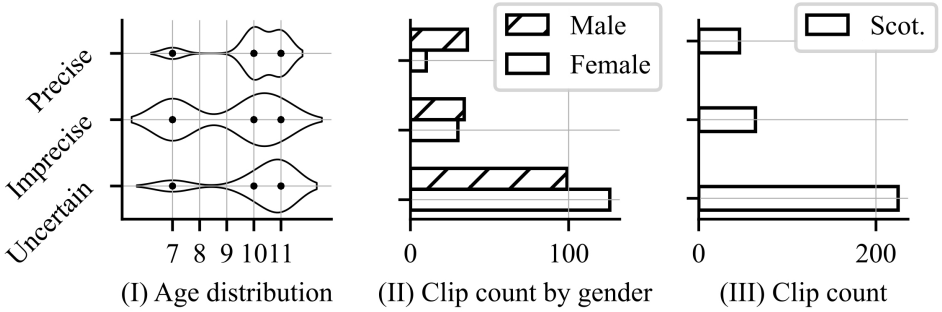
(a) The demographic distributions in the training dataset (China).



(b) The demographic distributions in the validation dataset (China).



(c) The demographic distributions in the test dataset (China).



(d) The demographic distributions in the test dataset (Scotland).

Figure 4.4: Demographic distributions in the A2M dataset.

boys and girls aged 7 to 11 years from two regions in China and one region in Scotland. Due to its greater relationship to the research team, Region A contributed a larger number of data samples.

Additionally, the A2M dataset was analyzed for data bias by investigating both representational and stereotypical biases, which are popular used in data bias studies (Dominguez-Catena et al. 2022). The representational bias is measured by the Normalized Standard Deviation (NSD) metric in Equation 4.2 (Dominguez-Catena et al. 2022), which evaluates the unequal representation of different demographic groups within the overall A2M dataset[5]. In the following equations, N denotes the number of demographic groups (e.g., male, female), $x_i^{(c)}$ is the count of instances of MMP c for group i, and $\mu^{(c)}$ is the mean count of MMP c across all groups. w_c represents the weight assigned to class c when calculating the overall NSD.

$$\mathrm{NSD}_c = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i^{(c)} - \mu^{(c)}\right)^2}}{\mu^{(c)} \cdot \sqrt{N-1}}, \mu^{(c)} = \frac{1}{N}\sum_{i=1}^{N} x_i^{(c)} \tag{4.1}$$

$$\mathrm{NSD} = \sum_c w_c \cdot \mathrm{NSD}_c \tag{4.2}$$

The stereotypical bias is measured by the Normalized Mutual Information (NMI) metric (Bouma 2009), assessing the variation in the demographic profile of each target class within the A2M[6]. The NMI is computed in Equation 4.3, where N denotes the set of demographic groups (e.g., male, female), and C represents the set of MMP classes (e.g., `precise`, `imprecise`, `uncertain`). $P(n, c)$ is the joint probability of observing demographic group n and MMP class c, while n and $P(c)$ are the marginal probabilities of the demographic groups and MMP classes, respectively.

$$\mathrm{NMI}(N, C) = -\frac{\sum_{c\in C}\sum_{n\in N} P(n,c) \ln \frac{P(n,c)}{P(n)P(c)}}{\sum_{c\in C}\sum_{n\in N} P(n,c) \ln P(n,c)} \tag{4.3}$$

The results of NSD and NMI are shown in Table 4.3. It is important to note that the NSD and NMI metrics do not share the same scales and are not directly comparable. The number in parentheses indicates the number of demographic groups considered: 5 ages, 2 genders, and 3 regions.

[5]The NSD value ranges from 0 to 1, where 0 indicates no bias and 1 indicates severe bias.

[6]The NMI value also ranges from 0 to 1, where 0 indicates no bias and 1 indicates severe bias.

Table 4.3: Representational and stereotypical bias metrics for the A2M

Dataset	Representational bias (NSD)			Stereotypical bias (NMI)		
	Age (5)	Gender (2)	Region (3)	Age (5)	Gender (2)	Region (3)
A2M	0.336	0.017	0.388	0.004	0.001	0.008

The results in Table 4.3 show that A2M exhibits minimal representational and stereotypical biases concerning gender. However, biases related to age and region were observed, and model bias is revisited during fine-tuning in Section 4.8.1.

Regarding data curation, the original raw data is restricted from sharing due to the General Data Protection Regulation (GDPR) as required by UK law (Great Britain 2018). However, the transparency guidelines outlined by Jo et al. (Jo & Gebru 2020) have been followed to make statistical data available to other researchers. Specifically, the user study prototype and statistical measurements of facial cues have been made publicly available on the GitHub site (github/affect2mmp 2025)[7]. This dataset, comprising 7,672 samples and 224 diverse features, provides a valuable resource for researchers to investigate facial cues during pupils' metacognitive monitoring and develop data-driven models.

4.6 Methods

The previous empirical studies used a conventional approach using the KMA to estimate MMP of pupils during gameplay (Kautzmann et al. 2016, Kautzmann & Jaques 2019). This section will introduce both the conventional and the proposed deep-learning approach based on facial expression interpretation.

4.6.1 Conventional Approach using the KMA

This section describes the conventional approach using KMA, which classifies MMP as either `imprecise` or `precise` (Kautzmann et al. 2016, Kautzmann & Jaques 2019).

[7]All statistical measurements of facial cues are stored in a table. In the table, each row represents one MMP clip, and columns are metadata (nation, gender, grade, age, task, attempt) and statistical measurements.

To estimate the MMP for the attempt p of task t, the conventional approach computes KMA. The KMA comprises four numbers (N_{c_1}, N_{c_2}, N_{c_3} and N_{c_4}), which correspond to the categories C_1 to C_4 illustrated in Figure 4.1. Each number N_{c_i} represents the number of times category c_i occurred in the previous k attempts. Using these four numbers, this approach calculates a coefficient (Equation 4.4) (Romesburg 2004), producing a KMA index that reflects the difference between pupils' JOC and their actual scores. The MMP is then estimated: if the KMA index exceeds a predefined empirical threshold θ, the MMP is deemed `imprecise`; otherwise, it is considered `precise`, see Section 4.7.5 for details of how to set θ.

The conventional approach uses k attempts, which are denoted as CA-k in the rest of the contents.

$$\text{KMA index} = 1 - \frac{(N_{c_1} + N_{c_4}) - (N_{c_2} + N_{c_3})}{N_{c_1} + N_{c_2} + N_{c_3} + N_{c_4}} \tag{4.4}$$

4.6.2 Deep-learning Approach: Meta-Facial Expression Interpreter

Deep learning has proven particularly effective for interpreting human facial expressions within video data. This research leverages this in Meta-Facial Expression Interpreter (M-FEI), which estimates pupils' MMP using video.

4.6.2.1 Inputs for M-FEI

Pupils' facial expressions were analyzed using two sets of features: the cropped facial region and facial cues. The cropped facial region captures the spatial aspects of facial expressions, providing pixel data from each frame that is suitable for image-processing neural networks. It preserves rich meta-information implicitly contained in facial images, such as gender[8], age, and emotional expressions, which can be inferred by deep learning models. Given that prior research has demonstrated the ability of deep learning networks to reliably recognize gender from facial images, this capacity has been widely adopted in facial analysis tasks (Li & Deng 2020). Consequently, the influence of gender on MMP estimation is implicitly encoded in the facial region.

The facial cues, consisting of action units (AUs), head pose, and gaze direction, see Figure 3.6 and Table 4.4, represent human-selected features related to cognitive

[8]This is an important factor for the correlation between facial cues and MMP, as illustrated in Chapter 3.

Table 4.4: Facial features extracted by OpenFace Baltrusaitis et al. (2018)

AUs	Head pose	Gaze direction
17 AUs in Figure 3.6	Location & rotation	Horizontal & vertical directions

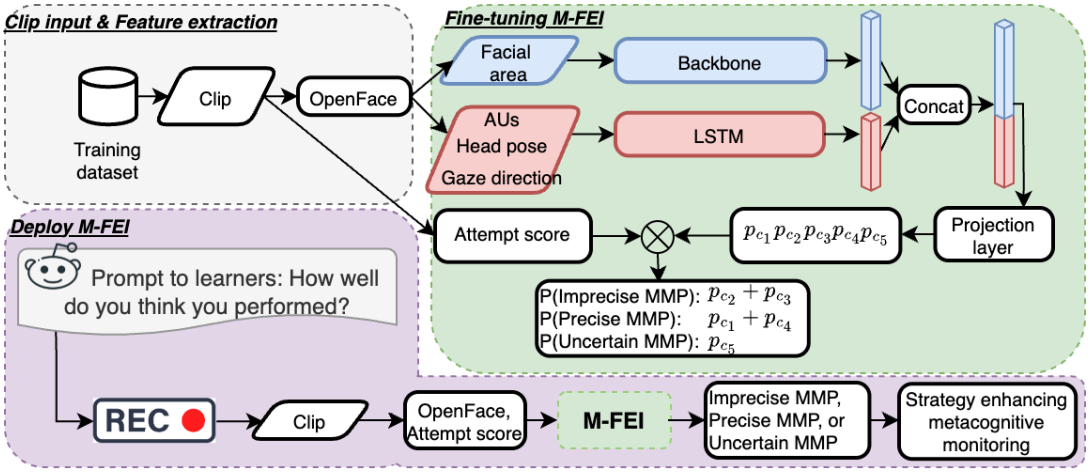


Figure 4.5: The M-FEI system follows a structured approach to process input clips. Grey region: data pre-processing and data input. Green region: M-FEI's network architecture. Purple: the workflow when deploying M-FEI.

processes (D'Mello et al. 2009, Eckstein et al. 2017, Pourmirzaei et al. 2023) and have been empirically validated to correlate with MMP in Chapter 3. The extraction of these facial cues was completed in Chapter 3.

By combining these two modalities, the deep learning model can effectively learn the complex relationships between implicit facial characteristics and explicit cognitive indicators, thereby improving the accuracy of MMP estimation. The following subsection explains how M-FEI processes these features.

4.6.2.2 M-FEI Initialization

The M-FEI was developed to process combined input features, as illustrated in the green section of Figure 4.5. Initially, using the OpenFace toolkit (Baltrusaitis et al. 2018), two key features were extracted from the primary input clips: cropped facial regions and facial cues, capturing temporal dynamics within each clip. These cropped facial areas are then analyzed through a pre-trained neural network[9] to interpret fa-

[9]Those pre-trained neural networks are available in (Hugging Face 2025).

Table 4.5: Pre-trained networks used by M-FEI. TimeSformer (K400/K600) (Bertasius et al. 2021) indicates that the model was trained and evaluated on Kinetics-400 (Kay et al. 2017)/Kinetics-600 (Long et al. 2020).

Network	**Pre-train dataset**	**Evaluation dataset**	**Acc.**
LTC (Donahue et al. 2015)	VGGFACE2	UCF101	91.7
TimeSformer (K400)	Kinetics-400	Kinetics-400	80.7
TimeSformer (K600)	Kinetics-600	Kinetics-600	82.2
X-CLIP (Ni et al. 2022)	Kinetics-400	Kinetics-400	81.1
SlowFast (Feichtenhofer et al. 2019)	Kinetics-600	Kinetics-600	81.8

cial expressions. Simultaneously, the facial cues are processed by a Long Short-Term Memory (LSTM) network[10] (Singh et al. 2023), which is widely used in prior research for handling temporal data due to its effectiveness in capturing sequential dependencies across time steps. The system then computes the estimated probabilities, p_{c_1} to p_{c_5}. Calibration of M-FEI's output is based on the pupil's AS, adjusting the probabilities accordingly: if the pupil's attempt is correct, p_{c_2} and p_{c_4} are set to zero (since they are not relevant); conversely, if the attempt is incorrect, p_{c_1} and p_{c_3} are zeroed[11].

For facial analysis within video frames, pre-trained networks listed in Table 4.5 were integrated to process cropped facial regions, as discussed in Section 4.3.4. These networks were selected based on their demonstrated accuracy in interpreting human behaviors in datasets such as VGGFACE2 (Massoli et al. 2020), UCF101 (Peng et al. 2018), Kinetics-400 (Kay et al. 2017), and Kinetics-600 (Carreira et al. 2018).

These all provide a comprehensive evaluation of MMP estimation. Additionally, for processing facial biometric markers, the long short-term memory (LSTM) model was adopted, due to its performance in analyzing learning-related facial expressions (Singh et al. 2023).

The total number of parameters in M-FEI is 120 million. The average fine-tuning

[10] An example of the LSTM network architecture is in Section 2.6.1.

[11] A masking operation was used to maintain the gradient flow during the network's propagation. AS was encoded into a mask code and used to perform element-wise multiplication with the vector $(p_{c_1}, p_{c_2}, p_{c_3}, p_{c_4}, p_{c_5})$. Specifically, the mask code for $AS = 1$ is $(1, 0, 1, 0, 1)$, and for $AS = 0$, it is $(0, 1, 0, 1, 1)$.

time per network (30 epochs) is 2,592 GPU hours (V100 GPU) on the Cirrus high-performance computing platform (EPCC 2025).

4.7 Experiments and Results

This section addresses two questions relevant to estimating pupils' MMP, as introduced in Section 4.2.

To address the RQ1.3, the performance of the conventional approach using KMA was evaluated on the A2M dataset.

To address the RQ1.4, the performance of M-FEI was first evaluated using the A2M dataset. It was then compared with the conventional approach in identifying `precise` and `imprecise` MMP. Inter-regional validation was subsequently conducted to assess the generalizability of M-FEI, and the risk of data bias, as discussed in Section 4.5.3, was also addressed.

In addition, the contribution of facial regions and facial cues to `imprecise` MMP was investigated. The findings were compared with prior research to highlight the relevance of specific facial cues as indicators of MMP.

4.7.1 Metrics

To quantitatively evaluate the performance of each approach, the Receiver Operating Characteristic Curve (ROC) was used as the evaluation metric (Kelleher & Hnin 2019). The Area Under the ROC (AUC) provides a measure of the model's ability to distinguish one category of MMP from another (Kelleher & Hnin 2019). A higher AUC value indicates better performance, with a value of 1.0 representing perfect discrimination.

When comparing M-FEI with conventional approaches across various binary classification scenarios, the confusion matrix and its derived metrics, True Positive Rate (TPR, also known as recall), False Positive Rate (FPR), and precision, were used (Stehman 1997). For each binary classification case, TPR (recall) measures the proportion of correctly identified instances among those defined as `positive`. FPR presents the proportion of incorrect `positive` identifications among the instances defined as `negative`. Precision indicates the accuracy of `positive` identifications, reflecting the proportion of true `positives` out of all `positive` identifications made.

Table 4.6: AUC values for the conventional approach when identifying `imprecise` MMP

CA	**A2M (exclude `uncertain`)**			
	Validation (China)	Test (China)	Test (Scotland)	Ave. (weighted)
CA-1	0.62	0.62	0.58	0.61
CA-2	0.65	0.63	0.62	0.63
CA-3	0.67	0.66	0.63	0.65
CA-4	0.68	0.68	0.65	0.67
CA-5	0.69	0.68	0.66	0.68
CA-6	0.69	0.68	0.66	0.68
CA-7	0.69	0.68	0.67	0.68
CA-8	0.69	0.69	0.67	0.68
CA-9	0.69	0.68	0.67	0.68

4.7.2 Overall Efficacy of Conventional Approach using KMA

In this section, the first question was addressed by estimating MMP using the conventional approach as explained in Section 4.6.1. The Meta-Brainhood game includes six different cognitive tasks. The performance of the conventional approach was assessed using up to nine previous attempts, with ten being the maximum number of attempts per task. If the number of completed attempts was fewer than k, the conventional approach relied on the available completed attempts for that task. It is important to note that the conventional approach is designed to identify only `precise` and `imprecise` MMP. Therefore, to ensure accurate evaluation, `uncertain` MMP instances were excluded from the validation and test datasets in this analysis.

Table 4.6 shows that as the number of attempts k increases, so does the averaged AUC of the conventional approach's estimation. The averaged AUC value rises to 0.68 when using four previous attempts and then plateaus at 0.69, showing no significant improvement thereafter. The conventional approach eventually yielded the highest average AUC value of 0.68 under stable conditions.

Table 4.7: AUC values for M-FEI when identifying `imprecise` MMP on validation dataset. Note: A for AUs, H for Head pose, and G for Gaze direction, see Table 4.4 for facial cues' details; Abbreviation w/ for 'with' and 'w/o' for without; The averaged AUC is weighted calculated by the ratio of size of validation dataset (excluding `uncertain` and including `uncertain`).

M-FEI		**A2M Validation**		
Pre-trained network	Features	excl. `uncertain`	incl. `uncertain`	Ave.
LTC	w/o AHG	0.83	0.76	0.80
	w/ AHG	0.81	0.77	0.79
TimeSformer (K400)	w/o AHG	0.84	0.78	0.81
	w/ AHG	0.84	0.76	0.80
TimeSformer (K600)	w/o AHG	0.80	0.77	0.79
	w/ AHG	0.84	0.78	0.81
X-CLIP	w/o AHG	0.78	0.73	0.76
	w/ AHG	0.81	0.74	0.78
SlowFast	w/o AHG	0.79	0.72	0.76
	w/ AHG	0.74	0.70	0.72
CA		excl. `uncertain`	-	Ave.
CA-8		0.69	-	0.69

4.7.3 Overall Efficacy of M-FEI

Deep learning neural networks were fine-tuned by combining features, as detailed in Section 4.6.2, using the A2M training dataset.

4.7.3.1 Overall Efficacy on Validation Dataset

The results of excluding `uncertain` MMP in Table 4.7 reveal that all M-FEI's AUC values surpass the random guess baseline of 0.5 and exceed the highest AUC value for the conventional approach. Notably, the TimeSformer models show superior performance, with the highest recorded AUC value of 0.84 in the validation, excluding uncertain results.

When the validation includes `uncertain` MMP, the results still indicate that all M-

FEI models surpass the random-guess baseline. However, the presence of `uncertain` MMP slightly reduces the AUC values across all model configurations. Despite this reduction, the TimeSformer-based M-FEI with the K400 again distinguishes itself, achieving the highest AUC value of 0.78. This consistent performance under both validation conditions underscores the better performance of TimeSformer-based models in understanding information in the video data in this research. This improvement can be attributed to their effective modeling of spatiotemporal dependencies and the ability to capture long-range interactions within facial expression sequences, which are critical for accurately estimating dynamic cognitive states.

4.7.3.2 Overall Efficacy on Test Dataset

To assess the generalizability and effectiveness of the fine-tuned M-FEI in real-world scenarios, its performance was evaluated on two test datasets comprising 16 Chinese pupils and 16 Scottish pupils whose facial expressions were not included in the training data.

The results of M-FEI on test datasets are presented in Table 4.8, and it shows varied performances across the pre-trained networks. In scenarios excluding uncertain MMP, the TimeSformer with AHG[12] notably had the highest performance in China's test dataset, achieving the highest recorded AUC of 0.74, which substantially surpasses the performance benchmarks set by other models. The TimeSformer (K600) with AHG also shows an outstanding 0.78 in Scotland. These results significantly surpass the baseline AUC of 0.5, suggesting a high level of predictive accuracy when identifying MMP.

Upon including `uncertain` MMP categories, which introduce the complexity and fuzzy cases, the TimeSformer (K400) with AHG continues to lead in performance, although with slight decreases. It records AUC values of 0.67 in China and 0.72 in Scotland. These results, while lower compared to the scenario of excluding `uncertain`, still outperform other tested models and configurations.

Compared to other models, such as LTC, X-CLIP, and SlowFast, whether with or without AHG, the TimeSformer (K400) consistently shows better performance across two test datasets.

[12]The TimeSformer pairing with AUs, head gestures, and gaze direction.

Table 4.8: AUC values for M-FEI when identifying `imprecise` MMP on test datasets. Note: A for AUs, H for Head pose, and G for Gaze direction, see Table 4.4 for facial cues' details; Abbreviation w/ for 'with' and 'w/o' for without; The averaged AUC is weighted calculated by the ratio of size of test datasets with excluding `uncertain` and including `uncertain`)

M-FEI		**A2M Test Datasets**				
Backbone	Features	excl. `uncertain`		incl. `uncertain`		
		China	Scotland	China	Scotland	Ave. (weighted)
LTC	w/o AHG	0.61	0.65	0.55	0.57	0.63
	w/ AHG	0.62	0.68	0.59	0.67	0.65
TimeSformer (K400)	w/o AHG	0.73	0.73	0.65	0.70	0.73
	w/ AHG	0.74	0.77	0.67	0.72	0.76
TimeSformer (K600)	w/o AHG	0.60	0.76	0.57	0.71	0.68
	w/ AHG	0.74	0.78	0.60	0.69	0.76
X-CLIP	w/o AHG	0.72	0.63	0.64	0.55	0.68
	w/ AHG	0.71	0.65	0.60	0.61	0.68
SlowFast	w/o AHG	0.71	0.70	0.61	0.59	0.71
	w/ AHG	0.70	0.62	0.61	0.53	0.66
CA		excl. `uncertain`		-		
		China	Scotland			Ave. (weighted)
CA-8		0.69	0.68	-	-	0.69

4.7.4 Threshold Calculations to Classify MMP

The overall performance of M-FEI was assessed using the AUC metric. As the AUC value reflects classifier performance across all possible thresholds, the optimal threshold for classifying MMP was determined using the Geometric Mean (GM) criterion (Espinosa et al. 2017).

M-FEI was evaluated in combination with TimeSformer (K400) and AHG, which yielded the best average AUC on the test datasets, for identifying `precise`, `imprecise`, and `uncertain` MMP. This configuration is denoted as M-FEI* for the remainder of this chapter.

In practical applications, determining the final MMP classification involves setting a specific threshold, θ. For instance, Kautzmann et al. (Kautzmann et al. 2016, Kautzmann & Jaques 2019) established a threshold of $\theta = 0.5$ for the KMA index, based on the results of a pre-test session conducted during their empirical evaluation study. The MMP is classified as `imprecise` if the KMA exceeds this threshold. For achieving the MMP classification by M-FEI*, three thresholds were established for M-FEI*'s three outputs $P(\texttt{precise})$, $P(\texttt{imprecise})$, and $P(\texttt{uncertain})$ based on the computation on three one-versus-rest (OVR) binary classifications (Psaltakis et al. 2024). The ROCs of the three OVR binary classifications are illustrated in Figure 4.6. The thresholds are computed using the GM criterion (Espinosa et al. 2017) as detailed in Equation 4.5, where X represents all clips in the validation dataset. The final selected category is the one with the highest score among those that exceed thresholds[13].

$$\theta_{GM} = \arg\max_{\theta}(\sqrt{TPR(X,\theta) * (1 - FPR(X,\theta))}) \tag{4.5}$$

Based on Equation 4.5, thresholds of 0.51, 0.27, and 0.22 were computed for the classification of `precise`, `imprecise`, and `uncertain` MMP, respectively. Additionally, M-FEI*'s performance at the computed thresholds is highlighted by A, and the CA-8's performance at the empirical threshold is indicated by B in Figure 4.6.

[13] In the results of this calculation, if no scores exceed their respective thresholds, the estimated MMP is classified as `uncertain`.

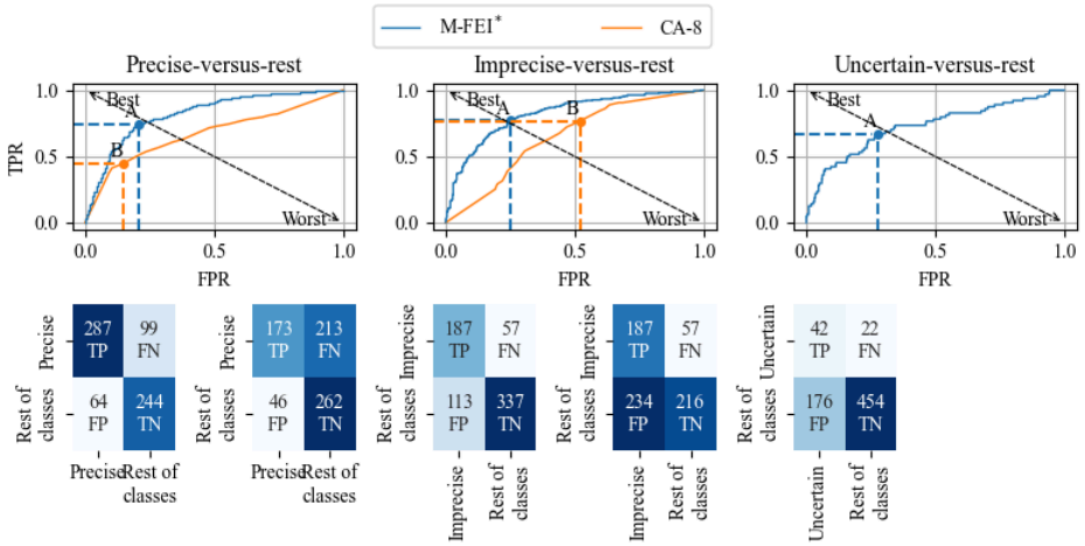


Figure 4.6: The Receiver Operating Characteristic Curves (ROC) for three one-versus-rest (OVR) binary classifications on the validation dataset, including `uncertain` MMP. The figure compares the performance of the M-FEI* in blue against the CA-8 in orange. This visualization is intended to illustrate the relative effectiveness of M-FEI* by juxtaposing it with the results from CA-8. Note: points A and B represent the network's performance at the computed thresholds; the left of the confusion matrices under each figure represents *A* on the blue line; the right one represents *B* on the orange line.

4.7.5 M-FEI* versus the Conventional Approach

To address the second research question, the performance of M-FEI* was first evaluated in identifying `precise`, `imprecise`, and `uncertain` MMP. Subsequently, M-FEI* was compared with CA-8, the CA-k configuration with the highest average AUC value of 0.68, under the condition of excluding `uncertain` MMP from the comparison. The comparison used the validation and two test datasets. The validation dataset contains 386 `precise`, 244 `imprecise`, and 64 `uncertain` MMP clips; the test dataset (China) contains 440 `precise`, 207 `imprecise` and 40 `uncertain` MMP clips; the test dataset (Scotland) contains 491 `precise`, 124 `imprecise` and 68 `uncertain` MMP clips.

Table 4.9: M-FEI* performance when identifying `Precise`, `Imprecise`, and `Uncertain` MMP

MMP	**A2M (including `uncertain`)**											
	Validation (China)				Test (China)				Test (Scotland)			
	Acc.	TPR	FPR	Pre.	Acc.	TPR	FPR	Pre.	Acc.	TPR	FPR	Pre.
`Precise`	0.77	0.74	0.21	0.82	0.64	0.62	0.31	0.78	0.72	0.73	0.31	0.86
`Imprecise`	0.76	0.77	0.25	0.62	0.70	0.59	0.26	0.50	0.75	0.66	0.24	0.38
`Uncertain`	0.71	0.66	0.28	0.19	0.55	0.40	0.44	0.05	0.70	0.35	0.27	0.13

The Table 4.9 presents the metrics of M-FEI* on A2M (including `uncertain`). As the test data from China and Scotland are unseen by M-FEI*, a commonly expected performance degradation on unseen populations is observed.

For the `precise` MMP category, M-FEI* achieves accuracies of 0.77 in China's validation and 0.64 and 0.72 in China's and Scotland's test datasets, respectively. The TPR (recall) is 0.74 in validation and 0.62 and 0.73 in the respective test datasets. The increase in FPR from 0.21 in validation to 0.31 and 0.31 in testing suggests an increasing difficulty in correctly identifying precise instances under varied test conditions. However, precision remains relatively stable, with figures of 0.82, 0.78, and 0.86 across the datasets.

For the `imprecise` MMP category, there is a slight decline in accuracy from 0.76 in validation to 0.70 and 0.75 in China's and Scotland's test datasets, respectively.

TPR (recall) also slightly decreases from 0.77 in validation to 0.59 and 0.66 in test datasets, and precision drops from 0.62 to 0.50 and 0.38. The FPR shows a stable level between 0.25 in validation and 0.26 and 0.24 in the test datasets. These shifts indicate a diminished effectiveness when identifying imprecise instances during testing, but a stable false positive rate in predictions.

In the `uncertain` MMP category, M-FEI* shows an accuracy of 0.71 in China's validation, which changes to 0.55 and 0.70 in China's and Scotland's test datasets, respectively. The TPR (recall) decreases significantly from 0.66 in validation to 0.40 and 0.35 in testing, while the FPR increases from 0.28 in validation to 0.44 and 0.27. Precision in `uncertain` estimations shows variability, with 0.19 in validation and dropping to 0.05 and 0.13 in testing. These results suggest that while M-FEI* is capable of identifying uncertain MMP, accurately distinguishing them remains challenging, particularly under varying test conditions. These findings indicate the need for a dedicated model specifically designed to handle `uncertain` MMP cases, enabling more accurate identification and interpretation of this challenging category.

It is observed that the M-FEI* model achieved higher accuracy and TPR on the Scotland test dataset compared to the China test dataset. This could be attributed to M-FEI* performing better in identifying `precise` MMP instances, which are more prevalent in the Scotland dataset (see Table 4.2). However, since the test sample size in each country includes only 16 pupils, it would be imprudent to draw strong conclusions from these results.

Table 4.10: Comparison of the performance of M-FEI* and CA-8 when identifying `imprecise` MMP

Method	**A2M (exclude `uncertain`)**											
	Validation (China)				Test (China)				Test (Scotland)			
	Acc.	TPR	FPR	Pre.	Acc.	TPR	FPR	Pre.	Acc.	TPR	FPR	Pre.
M-FEI*	0.78	0.77	0.21	0.70	0.70	0.59	0.24	0.54	0.83	0.64	0.12	0.56
CA-8	0.62	0.77	0.47	0.51	0.60	0.70	0.44	0.43	0.66	0.67	0.34	0.35

Table 4.10 reveals the comparison between two methods on A2M (excl. `uncertain`), on the validation, and two test datasets. For the `imprecise` category in the validation dataset, M-FEI* shows an accuracy of 0.78, showing a notable 26% improvement

over CA-8's 0.62. Especially, with a similar TPR (recall) of 77% compared to CA-8's 77%, M-FEI* significantly outperforms CA-8 in FPR at a lower level of 0.21, a 26% decrease. Also, M-FEI* achieves considerably higher precision at 0.70 versus 0.51 for CA-8. These results indicate that M-FEI* is more efficient at minimizing false alarms and more accurately identifying `imprecise` MMP instances than CA-8, while CA-8's overestimating pupils' lack of MMP could lead to detrimental responses, for example, over-frequent interventions.

In the test dataset for China, M-FEI* achieves an accuracy of 0.70, outperforming CA-8's 0.60. However, M-FEI* achieves a lower FPR at 0.24, significantly better than CA-8's 0.44, and maintains a higher precision of 0.54, compared to 0.43 for CA-8. The TPR (recall) for M-FEI* is 0.59, which is a bit lower than CA-8's 0.70, further confirming its efficiency in more accurately identifying imprecise MMP instances with fewer false positives.

As for the Scotland test dataset, M-FEI* continues to perform well with an accuracy of 0.83 versus 0.66 for CA-8. M-FEI* records a TPR (recall) of 0.64, similar to CA-8's 0.67. Furthermore, M-FEI* achieves a lower FPR of 0.12 compared to CA-8's 0.34, and a higher precision of 0.56 compared to 0.35. This performance indicates M-FEI*'s superior ability to reliably identify imprecise MMP instances while minimizing incorrect classifications.

A similar performance pattern is observed, with the M-FEI* model yielding higher accuracy and TPR on the Scotland dataset. While this trend may again relate to the prevalence of `precise` MMP instances in that dataset (see Table 4.2), the small sample size per country (16 pupils) cautions against over-interpretation.

Taken together, M-FEI* is capable of handling `uncertain` MMP, such as when pupils fluctuate between high and low confidence levels, although this remains a challenge, as reflected in the lower precision in experiments. Nonetheless, M-FEI* shows clear advantages in estimating the `precise` and `imprecise` MMP, offering greater accuracy and fewer false alarms compared to the conventional approach.

4.8 Additional Validations for M-FEI

Since metacognitive monitoring interventions in prior research have primarily been tailored to instances of `imprecise` MMP, where learners misjudge their own performance during self-reflection (Kautzmann et al. 2016, Kautzmann & Jaques 2019),

this part of the validations specifically focuses on evaluating the performance of M-FEI* in identifying `imprecise` MMP. Accurate detection of such cases is critical for delivering timely and adaptive support that helps learners recalibrate their confidence and improve MMP. This focus is particularly important in the second user study of this research, where metacognitive monitoring interventions are tailored based on the occurrence of imprecise MMP.

First, an evaluation examined model bias stemming from data bias in A2M. Next, an inter-regional test assessed M-FEI*'s generalizability across cultural contexts. Finally, an experiment explored which facial areas provide the most informative cues for MMP.

4.8.1 M-FEI* Model Bias

In Section 4.5.3, potential biases in the A2M dataset were identified. To investigate the impact of these biases on M-FEI*'s estimation, an experiment was conducted across age, gender, and region. Model bias was quantified using the Overall Disparity (OD) metric on the validation, test (China), and test (UK) datasets, the OD metric is measured by sample size-based weighted sum of intraclass disparity (see Equation 9 in (Dominguez-Catena et al. 2022)), where a value of 0 indicates no bias and a value of 1 indicates severe bias.

The OD was calculated by progressively training M-FEI* on random samples of the A2M dataset (ranging from 1% to 100%) and comparing the results with those from balanced A2M datasets[14]. As shown in Table 4.11, M-FEI* exhibits low bias across the analyzed categories. For age, the OD value decreased from 0.6 ± 0.1 to 0.2 ± 0.0 as larger samples were considered. For gender, the OD value declined from 0.7 ± 0.0 to 0.1 ± 0.1, and for region, it decreased from 0.7 ± 0.1 to 0.3 ± 0.0. These results indicate that the demographic biases identified in A2M influence M-FEI*'s estimations, and that increasing the sample size helps mitigate these model biases. Specifically, the model shows moderate biases related to age and regional attributes, and mild biases related to gender.

In addition, the model biases were validated by creating balanced datasets for age, region (A and B), and gender. However, compared with training on the imbalanced dataset (with the same size), balanced datasets for age and region do not effectively

[14] A balanced A2M dataset contains an equal number of samples for each MMP class across different demographic groups.

Table 4.11: The bias metric (Model Bias, OD) summary for M-FEI* is based on estimations using different sample sizes of the A2M dataset, with each sample size evaluated over 10 repetitions. Note: the balanced region configuration includes only data from Regions A and B.

A2M Training		**Size**	**Acc.**	**OD on Val. & Test (China & UK)**		
				Age (5)	Gender (2)	Region (3)
Original	1%	62	0.5 ± 0.3	0.6 ± 0.1	0.7 ± 0.0	0.7 ± 0.1
	2%	124	0.6 ± 0.1	0.5 ± 0.0	0.7 ± 0.1	0.6 ± 0.1
	4%	249	0.6 ± 0.1	0.5 ± 0.0	0.7 ± 0.1	0.6 ± 0.1
	8%	499	0.6 ± 0.0	0.3 ± 0.1	0.2 ± 0.1	0.3 ± 0.0
	16%	999	0.6 ± 0.0	0.3 ± 0.0	0.1 ± 0.0	0.3 ± 0.1
	43%	2705	0.7 ± 0.0	0.3 ± 0.0	0.1 ± 0.1	0.3 ± 0.1
	61%	3820	0.7 ± 0.0	0.3 ± 0.0	0.2 ± 0.1	0.3 ± 0.0
	98%	6140	0.7 ± 0.0	0.3 ± 0.0	0.2 ± 0.1	0.3 ± 0.0
	100%	6244	0.7 ± 0.0	0.3 ± 0.0	0.1 ± 0.1	0.3 ± 0.0
Balanced age	43%	2705	0.6 ± 0.0	0.3 ± 0.1	0.1 ± 0.0	0.3 ± 0.1
Balanced region	61%	3820	0.7 ± 0.0	0.3 ± 0.1	0.1 ± 0.1	0.3 ± 0.1
Balanced gender	98%	6140	0.7 ± 0.0	0.2 ± 0.0	0.1 ± 0.0	0.3 ± 0.0

mitigate model biases in these attributes. Only the balanced dataset for gender reduces OD from 0.2 ± 0.1 to 0.1 ± 0.0. Future research should investigate further strategies to mitigate model bias issues.

4.8.2 Inter-regional Validation: MMP Identification

The user study was conducted in two regions of China and one in Scotland. These diverse settings were selected to assess the generalization potential of the proposed method, considering regional variations as explained in Section 4.9.5. The demographics of pupils were shown in Figure 4.4.

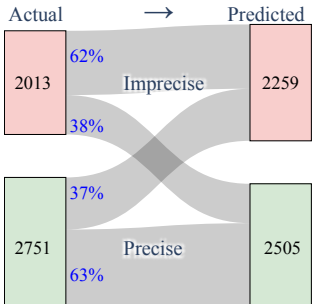
An inter-regional validation methodology was employed to evaluate M-FEI* in identifying `imprecise` MMP. Initially, M-FEI* was fine-tuned using data exclusively from either Region A or Region B and then tested on the alternate region. Subsequently, data from Regions A, B, and Scotland were combined for further evaluation.

The outcomes of this evaluation are presented in the flow chart of Figure 4.7, representing the MMP identification from clips of Regions A, B, and Scotland. It reveals that M-FEI* outperformed CA-8 in identifying `imprecise` MMP across all three regions while achieving far fewer false alarms. Despite identifying fewer `imprecise` instances in Region A, this is attributed to the limited training samples of Region B to fine-tune the model, as detailed in Figure 4.4. Based on the test results in Regions B and Scotland, M-FEI* achieved higher accuracy than CA-8 while significantly reducing the incidence of false alarms.

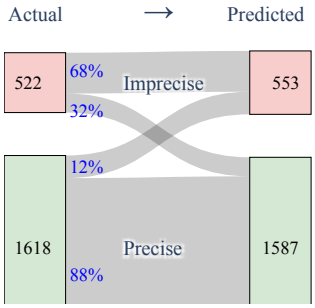
Thus, user study results indicate that M-FEI* consistently achieved higher precision, lower false alarms than CA-8 in inter-regional validation and attained higher TPR when sufficient sample data were available. However, further studies are necessary to verify whether these findings generalize to other contexts.

4.8.3 Informative Facial Features to the Best M-FEI

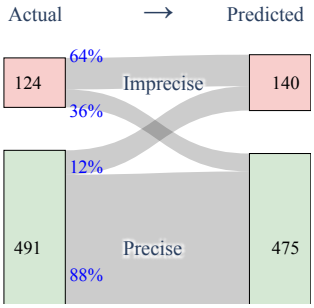
Given that M-FEI* outperforms conventional approaches in accuracy, including in the inter-region validation, a further analysis employed the Integrated Gradients approach (Sundararajan et al. 2017) to determine the sensitivity of M-FEI* to features made available by facial interpretation, enabling the selection of the most useful features. This quantified the attribute values of facial frame pixels and AUs values during inference over both validation and test datasets. This reveals the relative importance



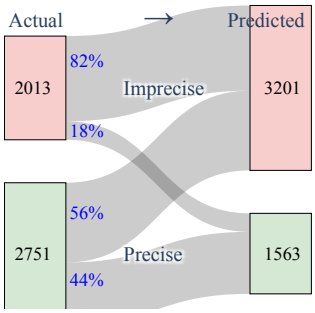
(a) Test M-FEI* in China (A)



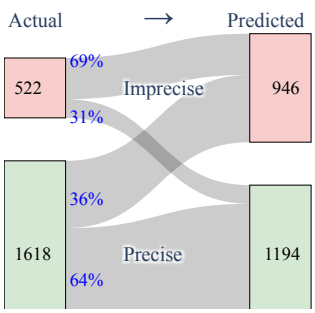
(b) Test M-FEI* in China (B)



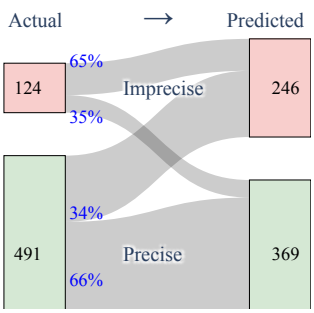
(c) Test M-FEI* in Scotland



(d) Test CA-8 in China (A)

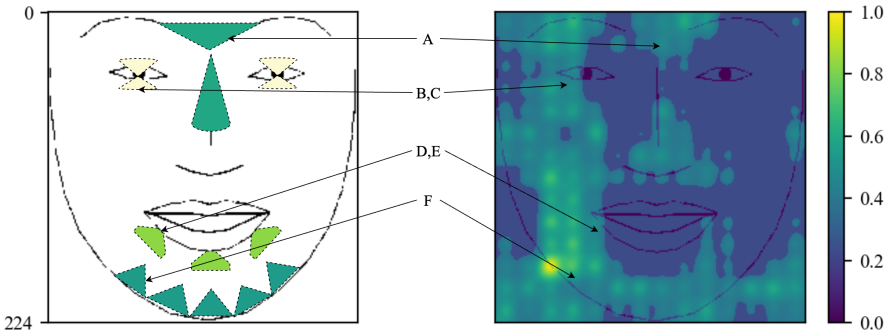


(e) Test CA-8 in China (B)

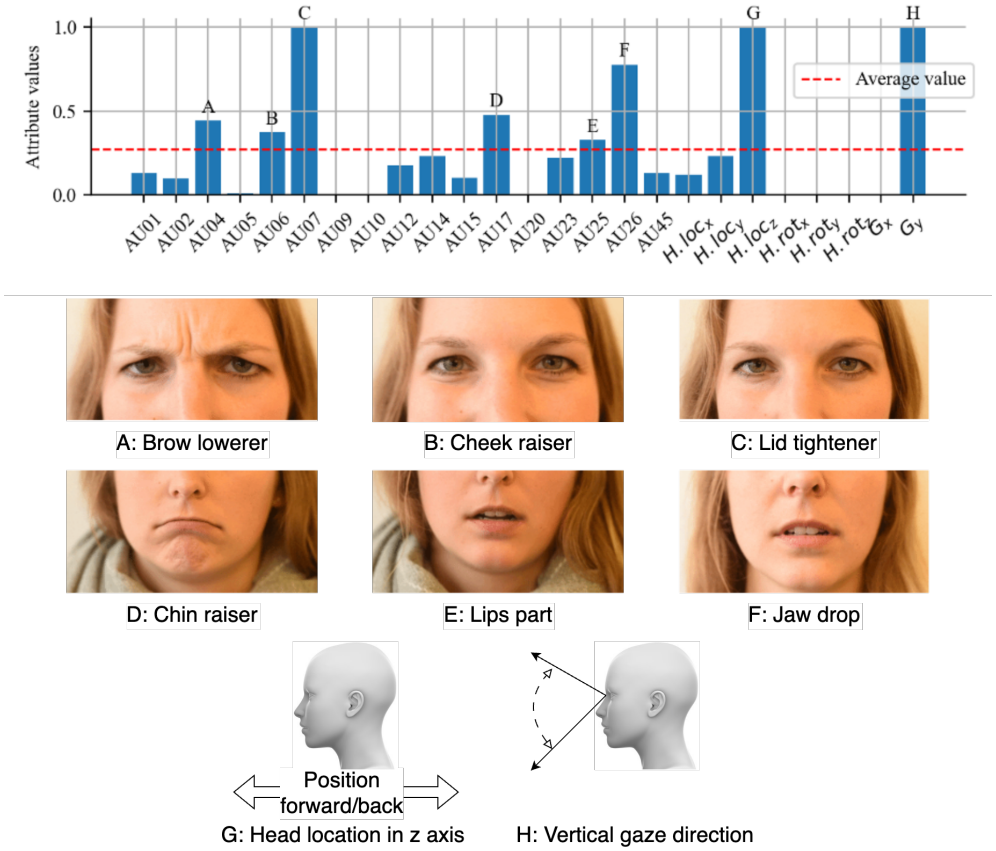


(f) Test CA-8 in Scotland

Figure 4.7: Inter-regional validation across China's Region A, B, and Scotland. The green and red indicators represent `precise` and `imprecise` MMP, respectively. In each sub-figure, the cubes on the left display the observed counts of MMP, while those on the right show the predicted outcomes. The blue numbers along the gray flows indicate the percentage of the clips classified into the target or error group.



(a) Facial template from OpenFace (left) and Heat map of standardized attribute values of facial movements in A2M (excl. `uncertain`) (right)



(b) Standardized attribute values of facial cues in A2M (excl. `uncertain`)

Figure 4.8: Facial features' attribute values derived by M-FEI*. Heat-map and bar chart height showing the sensitivity of M-FEI* output to each action in a facial image. The AUs' animations are from iMotions (2018).

of facial features for `imprecise` MMP identification.

Considering privacy, the results used a generic template face aligned with the facial region detected by OpenFace, as illustrated in Figure 4.8 (upper left). The standardized attribute values of facial features are visually represented as a heatmap spatially matching the abstract face (Figure 4.8, upper right), highlighting key areas such as the brow, eye, nose, cheek, and jaw, which are critical for estimating MMP. These regions are displayed with bright values denoting significance. Furthermore, the standardized attribute values of the AUs are depicted in a bar graph (Figure 4.8 (below)). Compared to the average attribute values, brow lowerer (AU04), cheek raiser (AU06), lid tightened (AU07), chin raiser (AU17), lips part (AU25), and jaw drop (AU26) emerged as particularly significant as impact when estimating `imprecise` MMP[15].

Using the Facial Action Coding System (FACS) (Ekman & Friesen 1978), significant AUs were mapped to specific facial areas, as shown in Figure 4.8 (upper left), where AU04 corresponds to area A, AU07 to area B, and so forth.

4.9 Discussion

This chapter aimed to investigate the feasibility of estimating pupils' MMP based on facial expressions. To achieve this objective, a large-scale user study was conducted to collect data, and M-FEI was fine-tuned to estimate MMP. The conventional approach using KMA for estimating pupils' MMP was assessed and used as the baseline for comparison. By comparing the performance of M-FEI* with that of the conventional approach, the feasibility of using M-FEI* to estimate pupils' MMP through the interpretation of facial expressions was demonstrated.

Findings of this chapter have the capability to enhance the mathematical learning context in any CBLE platform to tailor interventions adapted to MMP (shown in Chapter 5). Moreover, since our facial data reveals MMP in cognitive skills used in STEM[16], the M-FEI approach could be extended to broader similar learning contexts, such as physics, chemistry, and other subject areas (Brookman-Byrne et al. 2018, Träff

[15]In comparison with the analyses presented in Section 3.10, both sets of results indicate that brow lowering, cheek raising, lip parting, and vertical gaze direction are key facial cues associated with MMP.

[16]Science, technology, engineering, and mathematics (STEM) is an umbrella term used to group together the distinct but related technical disciplines of science, technology, engineering, and mathematics.

et al. 2019). This has the potential to influence the future development of CBLEs that adapt their metacognitive responses based on automatic analyses of learners' MMP. Findings in the results provide a foundation for future research on estimating MMP through the analysis of facial expressions.

4.9.1 Evaluations of the Conventional Approach

Recent studies of Kautzmann et al. have leveraged adaptive support to enhance pupils' MMP, particularly using the conventional approach that employs KMA (Kautzmann et al. 2016). There remains a notable gap in the literature concerning the accuracy of the conventional approach for estimating pupils' MMP within the context of CBLEs.

Validation in this chapter addresses this gap and confirms that the conventional approach achieves a stable accuracy for a diverse group of pupils, establishing its reliability. The research reported in this chapter also uncovered limitations of this approach, prompting the development of an automatic approach, M-FEI.

The conventional approach relies on aggregating a number of past JOC responses. This is problematic when insufficient historical data is available, e.g., a new task is presented or immediate repetition is not being used. The experiments undertaken in this chapter also shows that the conventional method plateaus when incorporating more than five previous attempts. Such a plateau may indicate a disconnect between past tasks (beyond the fifth prior attempt) and the current task, which may be exacerbated by variations in pupils' evolving knowledge and skills (Pintrich et al. 2000, Mihalca & Mengelkamp 2020). Such findings reinforce the need for improved MMP estimation strategies to better tailor metacognitive monitoring interventions.

Secondly, the results reported in this thesis support HCI research, showing that pupils face difficulties articulating their confidence (Harter 2012, Harris & Brown 2013, Amershi et al. 2014, Komatani & Nakano 2020, Lehnert 2024). In the A2M dataset, 9.5% of the responses were 'I do not know', highlighting a significant limitation on insisting that they evaluate themselves at one of the two extremes: 'Yes, I got that right' or 'No, I got that wrong'.

Lastly, the conventional approach can overestimate the frequency of `imprecise`, which can trigger excessive help, which may irritate pupils who do not require it (Flores & Lewis 2023). The CA-8's performance in both validation and test scenarios, in Table

4.10 and Figure 4.7, shows a consistent pattern of failures to accurately distinguish between `precise` and `imprecise` MMP. The FPR in the `imprecise` MMP, which reflects instances wrongly classified as `imprecise`, is alarmingly high (0.52 and 0.46 in validation and test, respectively). This indicates a systematic issue where nearly half of `precise` instances are misclassified as `imprecise`. This implies a significant likelihood of pupils receiving unnecessary interventions, thereby detracting from their learning experience and wasting resources.

4.9.2 Feasibility of Estimating MMP using Facial Expressions

Research exploring the relationship between facial expressions and metacognitive monitoring is still in its infancy, especially for pupils. Statistical evidence indicates that facial expressions are indicative of variations in MMP in undergraduate students (Taub et al. 2018, Taub & Azevedo 2018, Taub et al. 2021, Cloude et al. 2020). To our knowledge, the analysis undertaken in this chapter provides the first user study evidence supporting the use of predictive models to estimate pupils' MMP based on their facial expressions, extending prior work that primarily focused on correlation analyses. In this research, experimental results reveal that M-FEI is capable of identifying `precise`, `imprecise`, and `uncertain` MMP, although accurately identifying these `uncertain` MMP remains challenging. This capability is consistent across distinct regions. This shows the feasibility of using M-FEI to estimate MMP from facial expressions in a variety of contexts, at least for mathematical education, from the user study age group, using the given smart game.

The analysis undertaken here also assessed the impact of combining various facial biometric markers, such as AUs, head pose, and gaze direction, with a pre-trained model to recognize MMP more accurately. This showed that those combinations produce similar performance outcomes, suggesting that the cropped facial region used in M-FEI probably captured sufficient information.

Moreover, this research identified specific facial areas crucial for estimating pupils' MMP. Facial expressions, particularly those involving the movement of brows, lids, mouth, and chin, effectively indicate MMP.

M-FEI*'s capability suggests new avenues for investigating how metacognitive processes are reflected in facial expressions. Even though the current user study is mathematics-related, future studies should explore its applicability in other educa-

tional and training environments, or even in smart game contexts. This broadens the practical applications and deepens the understanding of metacognitive monitoring.

4.9.3 What are the Benefits of Adopting M-FEI in MMP Estimation?

The data leveraged in the research reported here was collected during mathematics-related cognitive tasks, indicating broad applicability in math and STEM learning. This suggests that M-FEI could prove valuable in several fields, including HCI, educational technology, and serious gaming.

When considering HCI, M-FEI estimates MMP using real-time spontaneous signals produced by learners to tailor metacognitive monitoring interventions. This eliminates the dependence on historical MMP data (Pintrich et al. 2000, Mihalca & Mengelkamp 2020). By using facial expressions, M-FEI enables personalized responses, which could adapt interactions to individual users' needs, significantly enhancing users' engagement and making interactions feel more intuitive and human-like (Behera et al. 2020, Baltaci & Gokcay 2016). M-FEI eliminates the need for learners to verbally express their confidence levels by automatically interpreting their facial expressions to identify metacognitive monitoring performance (MMP) as `precise`, `imprecise`, or `uncertain`. This aspect is particularly beneficial for young individuals with learning disabilities, who often find it challenging to articulate their feelings (Harter 2012, Harris & Brown 2013, Amershi et al. 2014, Komatani & Nakano 2020, Lehnert 2024). By providing an alternative method for assessing and supporting their cognitive state, M-FEI reduces reliance on verbal feedback and opens new avenues for adaptive educational technologies.

For educational technologies, M-FEI potentially enables educational platforms to tailor learning content to each learner's metacognitive state. Researchers can discover what the best response that an MMP estimate should trigger when a particular state is recognized for a particular age group learner tackling a particular task. Additionally, automatic MMP feedback could serve as a valuable tool for educators, particularly in MOOCs, where the educator-to-learner ratio is typically low. By providing real-time insights into learners' metacognitive states, such as when they are underconfident, overconfident, or struggling, M-FEI allows educators to monitor individual MMP trajectories without having to track each learner's performance manually

(Fauvel et al. 2018). This is especially important because different learners require different types and levels of support. For instance, a learner who consistently struggles with `imprecise` MMP may benefit from foundational guidance, while another who generally performs well but occasionally encounters setbacks might only require minimal or situational intervention. Addressing the diversity of learners' metacognitive needs, especially in large-scale educational settings, can be overwhelming for educators without technological support. By automating the detection of `imprecise` MMP, M-FEI helps reduce this burden, enabling teachers to provide targeted support at scale more efficiently.

For CBLEs, M-FEI's estimation of children's MMP could adjust content in the learning environment. For example, when a child displays underconfidence or overconfidence, the system could dynamically adjust its content or provide prompts to encourage reflection (Kautzmann et al. 2016). By responding to children's emotional and cognitive states in real-time, CBLEs can create more engaging and motivating experiences. Such responsiveness encourages sustained participation and enhances learning outcomes (Hamrouni & Bendella 2024). Particularly for children with anxiety or learning difficulties, a CBLE that automatically estimates MMP can provide targeted interventions. These interventions help build confidence and self-regulation skills in a low-pressure environment, making learning more accessible and effective (Ortegano & Ramírez 2019).

M-FEI stands to gain from advances in large vision models such as Llama (Touvron et al. 2023) and Qwen (Yang et al. 2024). Refinements in large vision models could significantly boost M-FEI's performance, enhancing its effectiveness and applicability in educational settings.

4.9.4 Threshold Selection and Performance Trade-offs when Identifying `Imprecise` MMP

Recognizing `imprecise` MMP is crucial in tailoring interventions. Kautzmann et al. set a low threshold for the estimated `imprecise` score, which allows most instances of `imprecise` MMP to be captured (Kautzmann et al. 2016, Kautzmann & Jaques 2019). However, this causes a high number of false alarms, with a large number of `precise` instances being misclassified as `imprecise`. M-FEI*, it is equipped with thresholds from Section 4.7.4 balancing discrimination among the `precise`,

`imprecise`, and `uncertain` categories, which, as a result, generate fewer false alarms.

It should be noted that in the context of educational settings, the trade-off between precision and recall in the M-FEI system designed to identify learners' `imprecise` is an affected state influencing both teaching strategies and student trust (Lindsley 1991). M-FEI interface provides customizable thresholds to education providers, so that they can adjust the balance between precision and recall.

In environments with sufficient educational resources, educators can choose to lower the threshold, prioritizing a higher recall to capture more instances of `imprecise`. Such a way leads to a decrease in precision, i.e., more false alarms. These can be mitigated through direct teacher-student communications (Kim 2024). Such mitigation allows for customization based on accurate assessments from educators and M-FEI.

In resource-constrained settings, high precision is crucial because the cost of misallocated interventions could significantly impact limited resources (Lindsley 1991). Educators may require a higher `imprecise` threshold. Such settings help in prioritizing students who are most likely to benefit, thereby optimizing the use of resources.

From the students' perspective, their tolerance for false alarms influences their acceptance and trust in the educational system (Oviatt et al. 2000). Teachers may seek to adjust the threshold according to their students' tolerance of false alarms (McAlenney & Coyne 2015). Adjustment is essential in maintaining student trust and ensuring that interventions are perceived as relevant and supportive rather than intrusive or misplaced.

Consequently, the balance between precision and recall should be strategically tailored to align with both resource availability and the psychological climate of the student body. Such a nuanced approach ensures that educational technologies enhance rather than detract from the learning experience.

4.9.5 Generalization Across Regional Variation

The user study reported in this chapter was conducted in China and Scotland. Populations in these regions differ in culture, traditions, economic status, and degree of urbanization. Additionally, distinct historical influences have shaped physical characteristics, including facial features, among these populations (Qiao et al. 2024). Leveraging these differences, additional experiments were then conducted to evalu-

ate how regional variations in facial expressions might affect M-FEI's performance in estimating MMP.

The inter-regional validation results in Section 4.8.2 reveal that the M-FEI* outperforms conventional approaches in MMP classification and performs with consistent accuracy. Such superior accuracy indicates that the model effectively generalizes across facial expression variations, which promises the potential for extending the application of M-FEI*.

A data curation plan was detailed in Section 4.5.3, which facilitates further collaborations to improve M-FEI's effectiveness in coping with diversity. The initial data repository has been settled in (github/affect2mmp 2025).

4.9.6 Selection of Facial Areas for Estimating MMP

As reported in Section 4.8.3, experiments were conducted to identify the specific facial areas most informative for MMP estimation. To the best of our knowledge, this is the first evidence to confirm that facial expressions, particularly those involving movements of the brows, lids, mouth, and chin, are effective in estimating pupils' MMP and better than the conventional approach. Prior research only shows that various facial expressions were indicative of precise and imprecise MMP (Taub et al. 2018, Taub & Azevedo 2018, Taub et al. 2021, Cloude et al. 2020).

Results reveal that spontaneous facial expressions can be effective in estimating MMP, opening new avenues for investigating how metacognitive processes are reflected through facial expressions in pupils.

4.10 Limitation and Future Work

The work and results detailed in this chapter have laid the foundations for several follow-up activities and studies, which are considered as further work.

Firstly, to improve data quality, a time-based filtering process was applied during pre-processing. Only video segments longer than 1.5 seconds were retained, as user study observations indicated that pupils who spent less time on the judgment of confidence task often responded without meaningful reflection, frequently engaging in blind clicking. While this filtering step helped reduce noise and enhance the reliability of the data, it also introduces a limitation to the research by potentially excluding

rapid but genuine responses, thereby affecting the generalizability of the findings.

Second, to our knowledge, the proposed approach is the first to estimate MMP automatically. While M-FEI alleviates learners from the challenge of articulating their confidence during metacognitive monitoring, there is currently no evidence to suggest that this convenience translates directly into improved learning outcomes. This lack of direct correlation to educational benefits leaves the broader impact of the approach somewhat uncertain. A second user evaluation study was conducted to specifically investigate the impact of M-FEI on learning outcomes in a mathematics context. The results of that user study will be reported in Chapter 5.

4.11 Chapter Summary

This chapter focused on evaluating the estimation of pupils' MMP using two distinct approaches. The primary contribution of this chapter is the development of a method for automatically estimating MMP by interpreting facial expressions, an ability traditionally demonstrated by effective teachers. This advancement opens important opportunities for supporting learners by integrating real-time MMP estimation into computer-based learning environments (CBLEs). This work makes four key contributions:

1. Developed the first dataset specifically designed for training deep learning neural networks to estimate MMP through facial expressions.

2. Revealed the performance and limitations of the conventional approach to estimating MMP.

3. Demonstrated the feasibility and benefits of using M-FEI to estimate MMP by interpreting facial expressions.

4. Developed a prototype system, M-FEI, that could be adopted and further enhanced by future research.

Chapter 5

User Study 2: Tailoring Metacognitive Interventions for Math-Solving by Responding to Facial Expressions

---

Only those who will risk going too far can possibly find out how far one can go.

— T. S. Eliot

One of the central goals of this research is to explore the RQ2.

In the previous chapter, a method was introduced to identify metacognitive monitoring performance (MMP) through facial expressions. However, a concern was raised that this automatic estimation might bypass learners' own confidence judgment processes, potentially affecting their engagement in self-regulation. This concern further strengthens the motivation to empirically validate whether such an approach can deliver educational benefits. To this end, the second user study was conducted to evaluate whether interventions tailored to MMP can positively impact pupils' performance in mathematics.

Building on the Meta-Facial Expression Interpreter (M-FEI) introduced in Chapter 4, which estimates learners' MMP using facial cues, this chapter develops an intelligent tutoring system (ITS) called Meta-Face Agent to provide tailored metacognitive interventions for learners. The Meta-Face Agent leverages a validated metacognitive intervention for MMP (Wood & Wood 1999, Kautzmann et al. 2016, Kautzmann & Jaques 2019), and it tailors the intervention to support a learner based on the esti-

mated MMP from M-FEI or a conventional approach, which is based on the knowledge monitoring assessment (KMA).

In this user evaluation study, 215 pupils participated. They were divided into two groups: (1) the 'M-FEI' group, which received the intervention tailored using the M-FEI method, (2) the 'KMA' group, which received the intervention using the conventional KMA approach. Findings indicated that pupils in the 'M-FEI' group not only showed greater improvements in learning outcomes following the intervention but also outperformed those in the 'KMA' group overall. This second user study, therefore, contributes to the advancement of tailoring learning systems by incorporating the interpretation of learners' facial responses, thereby enabling the development of more adaptive and effective computer-supported learning technologies.

5.1 Introduction for the Second User Study

As previously stated, self-regulated learning (SRL) is a multifaceted educational construct that encapsulates how learners orchestrate their cognitive, metacognitive, behavioral, motivational, and emotional processes to attain specific learning goals (Winne 2011). Central to the SRL framework is metacognitive monitoring, which enables learners to evaluate their progress, performance, and understanding of the content. This critical process includes activities such as verifying the accuracy of their solutions (Winne 2011). Empirical research, including studies by Isaacson et al. and Higgins et al. (Isaacson & Fujita 2006, Higgins et al. 2016), consistently illustrates a strong positive relationship between proficient metacognitive monitoring and enhanced exam scores. These insights emphasize the necessity of incorporating effective metacognitive support in the development of ITSs to improve learning outcomes.

Although metacognitive monitoring can significantly enhance learning outcomes, many learners do not engage in it naturally without targeted instructional prompts, such as guided questioning and detailed content explanations (Chi et al. 1989, Berardi-Coletta et al. 1995). Recognizing this limitation, much research has focused on developing support systems within ITSs to boost effective metacognitive monitoring (Grawemeyer et al. 2015, Kautzmann & Jaques 2019, Maras et al. 2019, Tisza et al. 2022, Cosentino et al. 2023). A common strategy is the implementation of text-based interventions to learners, which adapt to the learners' MMPs. ITSs such as Animated Pedagogical Agents (APA) (Kautzmann & Jaques 2019), Island Maths

Tutors (Grawemeyer et al. 2015), and Math Challenge (Maras et al. 2019) offer validated interventions that respond dynamically to a learner's performance. These systems provide more guidance when performance falls short and less when it meets or exceeds expectations (Wood & Wood 1999, Azevedo & Hadwin 2005).

Despite the potential benefits of these adaptive systems, delivering appropriate support relies on estimating the learner's MMP, which typically depends on self-reported confidence ratings. Based on the discussions in Chapter 4, such a self-articulation process can interrupt the flow of learning and may lack reliability (Ruan et al. 2025). Furthermore, accurately estimating MMP remains a challenge, especially for younger learners who find it difficult to articulate their confidence levels (Pintrich et al. 2000, Mihalca & Mengelkamp 2020). In response to these challenges, facial expression analysis has been investigated as an alternative method for assessing MMP. A novel technique, M-FEI, introduced in Chapter 4, uses spontaneous facial cues to more effectively identify MMP (Ruan et al. 2025).

The M-FEI technique changes the MMP estimation approach within CBLEs by eliminating the need for learners to articulate and report their confidence to the learning system (Ruan et al. 2025). This technique estimates MMP from learners' facial cues in response to self-reflective prompts such as, 'How well do you think you performed?' This method parallels traditional classroom interactions, where teachers evaluate students by direct questioning and assessing students' behaviors (Graesser 2020).

However, the effectiveness of M-FEI in delivering tangible learning outcomes within ITSs is still unclear. While M-FEI simplifies the MMP estimation process by removing the need for learners to articulate their confidence explicitly, this alteration could potentially reduce the depth of metacognitive engagement. Confidence judgments play a crucial role in self-assessment and are vital for developing an individual's awareness of their knowledge and skills (Pescetelli & Yeung 2021). These judgments of confidence (JOC) are instrumental in promoting reflective thinking, which is a core aspect of effective learning. A primary concern is whether M-FEI, by foregoing explicit self-articulation, might compromise the educational value of enhancing self-awareness. The work in this chapter aims to address these concerns by examining whether the advanced capabilities of M-FEI inadvertently undermine the learning benefits that come from fostering a deeper metacognitive awareness.

This chapter introduces the Meta-Face Agent, an ITS designed to enhance learn-

ers' MMP while solving math exercises. The Meta-Face Agent tailors a validated metacognitive intervention by integrating two approaches of estimating MMP (Wood & Wood 1999, Kautzmann et al. 2016, Kautzmann & Jaques 2019): the KMA-based method and the newly developed M-FEI. This dual approach enables investigation into the effects of tailoring interventions based on different MMP estimation methods on learners' performance. It also offers flexibility for future educators using the Meta-Face Agent to select the most appropriate method according to specific teaching contexts. The conventional MMP estimation approach requires the learners to explicitly state their confidence in their answers, typically choosing between binary options like 'My answer is right' or 'My answer is wrong.' Conversely, the M-FEI-based approach estimates MMP by prompting learners with the question, 'How well do you think you performed?'. This prompt appears after each response, and the M-FEI system analyzes the learners' facial expressions to infer their metacognitive monitoring performance.

The Meta-Face Agent delivers the validated metacognitive interventions (Wood & Wood 1999, Kautzmann et al. 2016, Kautzmann & Jaques 2019) through a series of text-based prompts, designed to enhance learners' MMP and improve overall learning outcomes. These prompts are structured into three progressive levels, as established by previous research (Kautzmann et al. 2016, Kautzmann & Jaques 2019): The first level encourages reflection on prior knowledge relevant to the current task. The second level prompts learners to recall pertinent knowledge taught in the classroom. The third level guides them through the necessary problem-solving steps for similar tasks (Wood & Wood 1999). This prompt-based intervention is activated only when the estimated MMP from the learner indicates an `imprecise` MMP.

The design and deployment of the Meta-Face Agent in the user evaluation study aimed to address RQ2 from Chapter 1.

A user evaluation study was conducted with 215 pupils from grades 1 through 6, aged 7 to 12[1], to assess math-solving outcomes supported by the Meta-Face Agent.

Pupils in the study were divided into two groups, each getting interventions based on a different MMP estimation approach in the Meta-Face Agent: the M-FEI-based or the KMA-based. Participants answered a series of math questions, enabling us to address RQ2.1 and RQ2.2 by comparing mathematical learning outcomes between the two groups.

[1]For clarity, 66 pupils were in user study 1.

To be precise, the observed improvements in mathematical learning outcomes may be partly attributed to a practice effect, as learners were allowed to revise their answers during the text-based intervention. An additional test was conducted to analyze pupils' learning outcomes when they meet similar math questions for the second time. This test is used to evaluate the extent to which the practice effect influenced these outcomes.

The findings indicate that both interventions, tailored based on M-FEI and the KMA-based MMP estimation approach, contributed to improved math-solving outcomes in the initial test stage. Notably, the improvement observed in the 'M-FEI' group was more substantial.

5.2 Related Work for Intelligent Tutor Systems with Adaptive Metacognitive Support

This related work section begins by reviewing the critical role of metacognitive skills in learning. It then examines prior ITSs designed to support metacognitive monitoring. Following that, it discusses the benefits of incorporating AI techniques into ITSs to enable adaptive support, as well as the limitations of metacognitive support provided in existing systems.

5.2.1 Relationship between Metacognitive Skills and Learning Outcomes

Research in the past decade strongly indicates that students' metacognitive abilities are positively associated with their learning outcomes. A meta-analysis done by Ohtani & Hisasaka (2018) reveals that metacognitive skills, while only moderately correlated with grades on average, remain a significant independent predictor of academic achievement. Domain-specific studies likewise report that learners with better metacognitive skills tend to attain deeper understanding. For example, a recent experimental study in science education observed a positive correlation between high school students' metacognitive skill level and their conceptual understanding of science topics, especially when a self-organized learning environment pedagogy was used (Tsamago & Bayaga 2024). In higher education, Halmo et al. similarly report that first-year students who excel in metacognitive skills, including planning, monitor-

ing, and evaluating their problem-solving, achieve better academic outcomes (Halmo et al. 2024). These empirical studies, spanning correlational analyses, classroom interventions, and even think-aloud protocols, converge on the idea that students who actively reflect on and control their learning process perform better. The evidence has encouraged educators to incorporate metacognitive strategy training into curricula, as improving students' metacognitive skills can translate into improved test performance, deeper conceptual mastery, and more efficient learning (Halmo et al. 2024). In summary, a robust body of recent work underscores that enhancing metacognition is not just an abstract goal; it has measurable impacts on learning outcomes across age groups and disciplines.

5.2.2 Intelligent Tutoring Systems Supporting Metacognition

Traditional ITSs focused on step-by-step guidelines for problem-solving (e.g., in math or physics), but new systems integrate metacognitive interventions like prompts to self-explain, hints to encourage reflection, and tools for students to plan and assess their understanding. For instance, MetaTutor is an ITS that has been at the forefront of metacognitive support research (Azevedo et al. 2009, D'Mello & Graesser 2013). It is a hypermedia learning system for human biology where students learn by exploring content with the guidance of animated pedagogical agents. These agents act as external metacognitive regulators, prompting the learner to set goals, monitor progress, and adjust strategies, which are essential to metacognitive monitoring in SRL (Azevedo et al. 2022). Over a decade of studies with MetaTutor compared conditions with adaptive metacognitive scaffolding versus no scaffolding, consistently finding that students who received the tailored prompts to plan, monitor, and self-evaluate showed higher learning gains and better regulation behaviors (Azevedo et al. 2022). Another example is an experimental logic tutoring system studied by Abdelshiheed, Hostetter, Shabrina, Barnes & Chi (2023), which taught students when to switch problem-solving strategies. In that study, students initially defaulted to a single strategy; the ITS then provided different types of interventions (e.g., worked examples, nudging prompts) to encourage more strategic metacognitive behavior. The results showed that a gentle nudge strategy, where the tutor prompted the student at opportune moments to consider an alternative approach, led to significant improvement. This demonstrates how ITSs can embed targeted metacognitive interventions

within problem-solving practice.

In text-based learning environments, conversational ITSs like AutoTutor (D'Mello & Graesser 2013) have incorporated metacognitive dialogue moves. For example, it asks the learner, 'Why do you think that answer is correct?' or 'How confident are you?' to prompt reflection. These interventions are analogous to a human tutor's probing questions and have been shown to deepen comprehension and improve students' calibration of their understanding.

5.2.3 AI-driven Adaptive Support to Metacognitive Monitoring

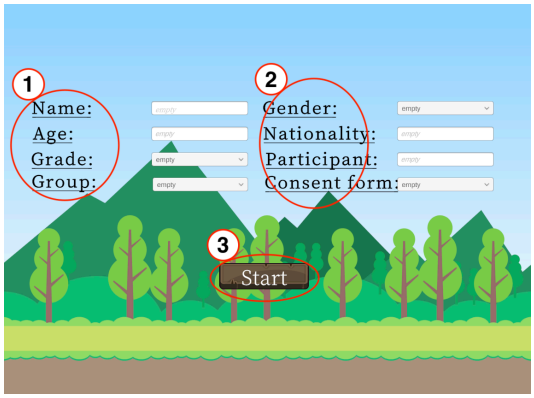
One of the biggest advantages of incorporating AI into the learning system is the ability to adapt support to individual needs in real-time. Learning support, like generic worksheets or static prompts, is the same for everyone and cannot respond if a student misunderstands a concept or skips an important reflection step. AI-driven systems can monitor learners' actions and performance, and then tailor metacognitive interventions. It can decide when to prompt, what hint or question to offer, and how to personalize it, thereby closely mirroring the kind of guidance a skilled human tutor might provide (D'Mello & Graesser 2013, Graesser 2020). Recent research provides concrete examples of the benefits of adaptation to interventions. Wang et al. (2023) studied medical students using an ITS called BioWorld, which adapts targeted prompts based on students' behaviors. The results found that students who engaged more with the metacognitive hints became better calibrated. Their confidence judgments more accurately reflected their actual performance, and they solved problems more efficiently. Another compelling example comes from the use of machine learning to optimize metacognitive interventions. Abdelshiheed, Hostetter, Barnes & Chi (2023) deployed a deep reinforcement learning (DRL) agent to decide when and how to coach students on strategy use across two different tutors. In their classroom experiments, one group of students received fixed, rule-based strategy prompts (based on a simplistic classifier of their metacognitive skill level), while another group received AI-adaptive prompts tuned by the DRL agent. The adaptive system decides to intervene by observing student behavior and outcome patterns. The results were striking: the DRL-based adaptive interventions closed the metacognitive skills gap between students of different initial abilities, whereas the static group-based prompts

only helped students who already had some strategic knowledge. Moreover, students taught with the adaptive AI support showed better transfer of learning. In weeks later, when all students moved to a new tutor in a related domain without any prompts, those who had experienced the adaptive metacognitive coaching performed significantly better than their peers.

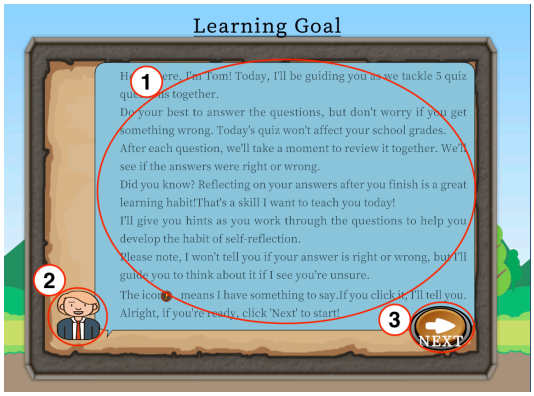
5.2.4 Challenges of Metacognitive Support in ITSs

Despite the promising results of metacognitive support, significant challenges persist in how these systems compare to the nuanced guidance offered by human teachers. Historically, traditional computer-based support had limited insight into learners' thought processes. For example, prior works adopted a KMA-based approach, which is outlined in Chapter 4. Those works typically provided hints or feedback based solely on the correctness of answers, without considering how students arrived at those answers or their emotional responses to the material (Kautzmann et al. 2016, Guo 2020). Modern AI-based ITSs have improved on this by tracking learners' response times, hint usage patterns, confidence ratings, and other metrics. For example, the BioWorld study required students to explicitly report their confidence in their diagnostic decisions, enabling the system to assess their MMP (Wang et al. 2023). However, unlike a human teacher who can intuitively sense confusion or disengagement in the classroom and pose spontaneous metacognitive questions, ITSs struggle to derive MMP from indirect proxies like spontaneous responses (such as facial cues) or errors in work (Azevedo et al. 2022). This situation underscores the need for new technologies that can gather and interpret multimodal data more akin to the way a human might naturally infer learners' metacognitive states.

Although accurately gauging a learner's metacognitive state remains a significant challenge, the approach presented in Chapter 4 offers a promising direction. By analyzing learners' emotional expressions, this method enables the automatic tailoring of support based on their momentary cognitive states. Importantly, it helps maintain the learning flow and demonstrates greater accuracy than the KMA-based MMP estimation approach. In this way, M-FEI equips ITSs to provide support in a manner more akin to a human teacher, responding intuitively to signs of confusion or disengagement as they naturally emerge in the learning process.



(a) Login page in Mate-Face Agent. 1 (from up to down): Name, age, grade, group; 2 (from up to down): Gender, nationality, participant ID, consent form; 3: Start.



(b) Learning introduction in Meta-Face Agent. 1: Introduction of this study; 2: The Meta-Face Agent's avatar in the ITS. 3: Next.

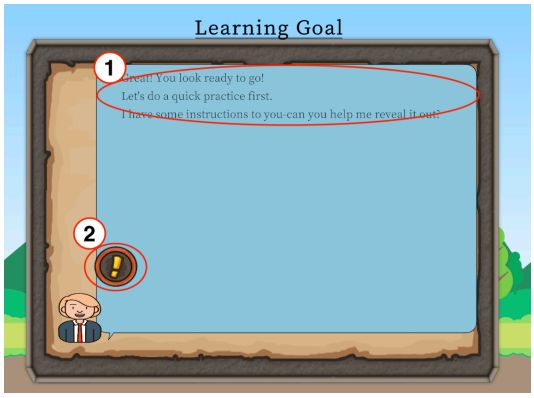
Figure 5.1: Meta-Face Agent interface overview: Login and learning goal

5.3 Meta-Face Agent

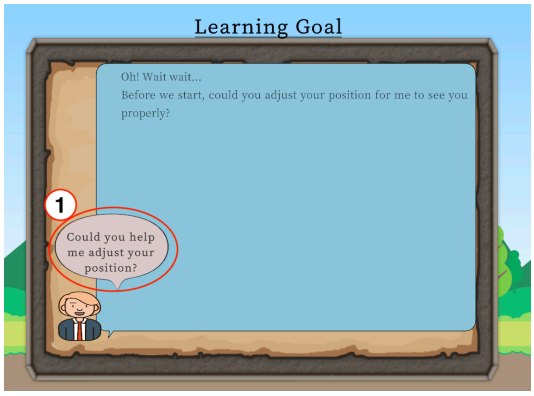
This section introduces the setup of the Meta-Face Agent, including its user interface, agent design, and overall workflow. It then describes how the agent estimates MMP, including the KMA-based approach and the M-FEI-based approach. Finally, it details the tailored metacognitive monitoring intervention and the manner in which they are delivered.

5.3.1 General Setting

The Meta-Face Agent is an innovative learning system featuring a half-body character (i.e., an Avatar) that remains visible on the graphical interface throughout the pupils' interaction, as illustrated in Figure 5.1. The UI features are based on the APA, a validated, well-established, CBLE (Kautzmann et al. 2016, Kautzmann & Jaques 2019). At the beginning of the study, the agent will introduce two practice sessions to learners about how to check metacognitive prompts from the agent and how to adjust their position, ensuring the face area is captured in the webcam scope, see Figure 5.2. This character is designed with a friendly, smiling facial expression to create a welcoming and engaging learning environment. Textual messages from the character are presented in speech bubbles, enhancing readability and keeping the interface clean



(a) Check agent's prompts. 1: A practice to check the agent's instructions. 2: Icon for instruction.



(b) Adjust position. 1: A practice to adjust the user's sitting position.

Figure 5.2: Meta-Face Agent interface overview: Practice interaction with agent

and well-organized. Additionally, these messages are audibly delivered using OpenAI's synthesized human voice technology (OpenAI 2025), which provides a natural and accessible user experience.

The Meta-Face Agent leverages the metacognitive intervention validated by APA (Kautzmann et al. 2016, Kautzmann & Jaques 2019), which is used to enhance learners' MMP. The designed intervention is delivered by providing prompts and text-based messages to learners to actively encourage them to reflect on their knowledge and understanding within a subject area. Tailored specifically for mathematics, the prompts within the Meta-Face Agent were structured into three distinct levels to progressively guide and enhance learners' MMP, following the framework proposed by Wood & Wood (1999):

Lv1. Encourage learners to identify what the problem is asking for. For example, 'Wait, can you reflect a bit on what the question is asking? Take a moment to think, and then go forward if you are ready.'

Lv2. Encourage learners to think about their previous knowledge. For example, 'Take a moment to think—do you remember the properties of a square that we discussed in class? And what about the knowledge of a circle's diameter and radius?'

Lv3. Encourage learners to think about the steps of solving similar questions that were solved before. For example, 'Let's recall if we have done similar exercises

using a compass before. Think about how to use a compass to draw a circle. Remember the steps to solve similar problems. There's a similar problem on page 2 of the workbook we prepared for you. Review the steps to solve this type of problem.'

In the intervention session, the frequency and level of the prompts delivered by the Meta-Face Agent are carefully adapted to meet individual student needs. The specific workflow of the intervention, along with how the intervention is tailored based on the MMP estimation approaches, is detailed in Section 5.3.3.

5.3.2 Estimation of MMP

5.3.2.1 KMA-based Kernel

Metacognitive monitoring involves assessing one's own knowledge, learning, and cognitive states either during or after task performance. This cognitive process includes making judgments about confidence in one's understanding, accuracy of recall, or awareness of comprehension (Flavell 1979). To estimate MMP, researchers often rely on a JOC scale post-task. For instance, subjects might rate their confidence in the correctness of their answers on a scale from 0 (completely unsure) to 1 (completely sure) (Koriat 1981). The agreement between these JOC ratings and actual performance provides a quantifiable measure of learners' MMP.

As a reminder of the more detailed discussion in Section 4.6.1, the MMP estimated by the KMA-based kernel is considered by the KMA index in Equation 5.1. The definition of c_1 to c_4 is depicted in Figure 5.3. c_1 to c_4 represent different cases of the learner's judgment:

'+ +' indicates that the learner judged the answer as 'correct', and indeed, the answer was correct.

'+ -' applies when the learner judged the answer as 'correct' but the answer was incorrect.

'- +' refers to situations where the learner judged the answer as 'wrong', yet the answer was correct.

'- -' is used when the pupil judged the answer as 'wrong' and the attempt was also solved incorrectly.

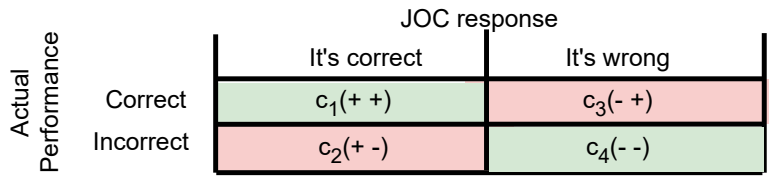


Figure 5.3: Four classes of metacognitive monitoring performance: The matrix of task performance and JOC response

Using the occurrences of these four cases, the KMA-based kernel calculates the KMA index as an estimation of MMP.

$$\text{KMA index} = 1 - \frac{(N_{c_1} + N_{c_4}) - (N_{c_2} + N_{c_3})}{N_{c_1} + N_{c_2} + N_{c_3} + N_{c_4}} \tag{5.1}$$

5.3.2.2 M-FEI-based Kernel

The novel deep learning-based neural network, M-FEI (described in Chapter 4), was introduced to estimate learners' MMP from spontaneous facial expressions. The operational framework of M-FEI within the Meta-Face Agent is illustrated in Figure 5.4. This system processes video clips of facial expressions and outputs the probability that the learner's MMP is `precise`, `imprecise`, or `uncertain`. The neural network's ability to analyze subtle facial expressions allows for a nuanced understanding of learners' metacognitive states, potentially used to tailor the intervention, enhancing the accuracy of MMP in educational settings.

Here, we provide a resource consumption overview of running the M-FEI kernel. On a typical laptop equipped with an Apple M1 CPU and 16 GB of system memory, the mean inference time was approximately 3.17 seconds per MMP estimation. The peak CPU memory usage during estimation was 231 MB, indicating that the model can be run efficiently without exhausting system resources. The total number of floating-point operations required for an estimation was estimated at 3.9 trillion (FLOPs). These results suggest that M-FEI can perform real-time or near–real-time inference on widely available consumer-grade hardware without the need for dedicated GPUs or remote servers.

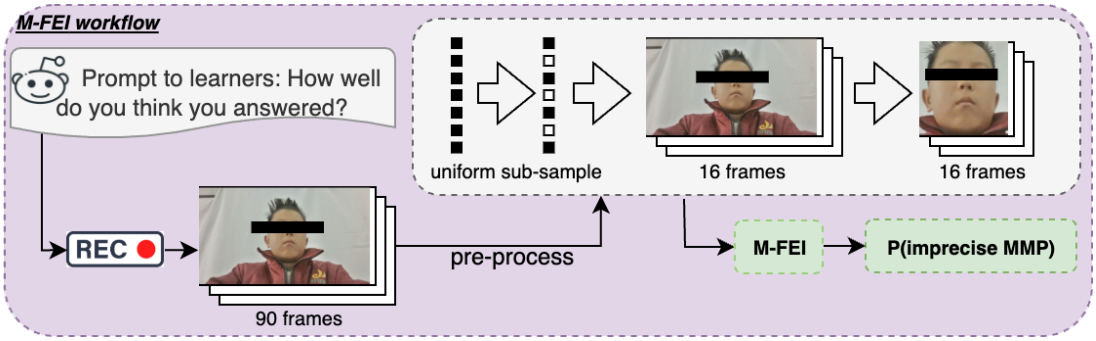


Figure 5.4: M-FEI kernel's workflow for MMP estimation: Following a prompt to the learner, facial expression frames are sampled and transformed. These processed frames are then analyzed using the deep-learning model, M-FEI, to estimate the learner's MMP.

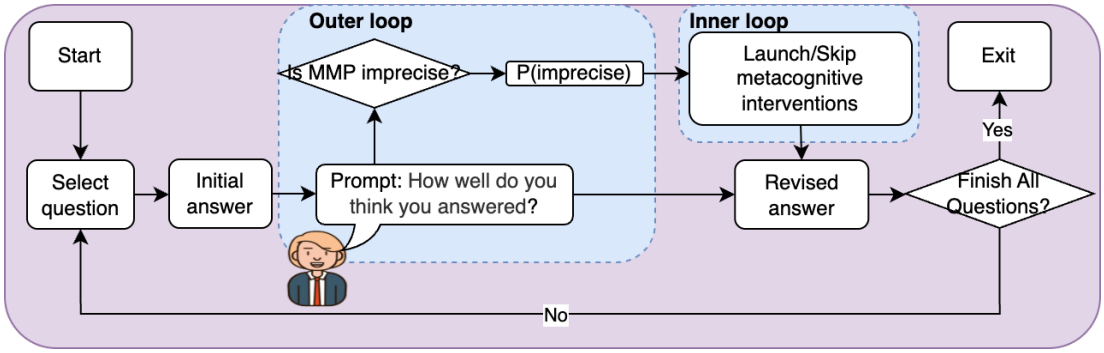


Figure 5.5: Outer and Inner loop in Meta-Face Agent.

5.3.3 Adaptation of the Metacognitive Intervention in Meta-Face Agent

Previous research has demonstrated the efficacy of support systems within ITSs to enhance learners' learning (Azevedo & Hadwin 2005), particularly through metacognitive interventions aimed at improving learners' MMP (Kautzmann et al. 2016). Inspired by the successes of APA (Kautzmann et al. 2016), the Meta-Face Agent incorporates a dual-loop system, comprising both an outer loop and an inner loop, to tailor the metacognitive intervention effectively, see Figure 5.5. The subsequent paragraphs detail the specific adaptation mechanisms employed by the Meta-Face Agent.

5.3.3.1 Outer Loop

This mechanism is pivotal in activating the metacognitive intervention managed by the inner loop of the Meta-Face Agent and operates continuously as learners reflect on their answers. Before learners transition to the next question, the outer loop poses the question, 'How well do you think you answered?', see Figure 5.6a. Based on the response, one of the MMP estimation kernels, either KMA-based or M-FEI-based, is selected. The likelihood of triggering the metacognitive intervention increases with the probability of an `imprecise` MMP assessment. By aligning the frequency of giving intervention with each learner's MMP, this mechanism ensures that support is tailored and timely.

5.3.3.2 Inner Loop

This mechanism is responsible for delivering the validated prompt-based metacognitive intervention within the Meta-Face Agent. It uses metacognitive monitoring prompts previously validated in studies by Kautzmann et al. (Kautzmann et al. 2016, Kautzmann & Jaques 2019). These prompts are crafted to encourage learners to reflect on the current questions and to recall their knowledge from similar previously solved questions. Examples of such metacognitive prompts are showcased in Figure 5.7.

Depending on the result from the outer loop, the Meta-Face Agent either activates the appropriate metacognitive intervention through the inner loop or skips this step (if the estimated MMP is `precise`, i.e., no intervention is deemed necessary).

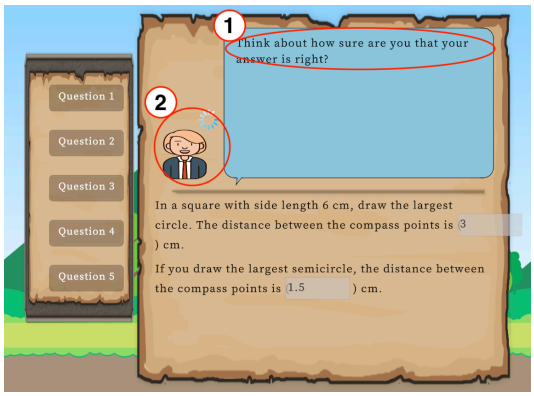
The selected prompts are displayed in a bubble box adjacent to the Agent's character and are voiced using OpenAI's synthesized human voice, enhancing the interactive experience. After engaging with the prompts, learners are prompted to reflect on their answers and proceed to the next question, see Figure 5.6b.

The Meta-Face Agent, equipped with the KMA-based kernel, analyzes learners' recent history of MMP to compute a KMA index ranging from 0 to 2[2]. A threshold was established in the study, classifying any KMA index greater than 0.5 as `imprecise`

[2]The KMA index is defined as:

$$\text{KMA index} = 1 - \frac{(N_{c1} + N_{c4}) - (N_{c2} + N_{c3})}{N_{c1} + N_{c2} + N_{c3} + N_{c4}} = \frac{2(N_{c2} + N_{c3})}{N_{c1} + N_{c2} + N_{c3} + N_{c4}}.$$

Since all $N_{ci} \geq 0$, the index is bounded within the range $[0, 2]$, where it reaches 0 when $N_{c2} + N_{c3} = 0$ and 2 when $N_{c1} + N_{c4} = 0$.



(a) Metacognitive stimuli. 1: 'How well do you think you performed?'; 2: GIF while M-FEI is running.

(b) Submit answer and move to next question. 1: Instruction to select the next question; 2: Completed questions.

Figure 5.6: Start and end of each metacognitive prompt in Meta-Face Agent

MMP. The inner loop of the Meta-Face Agent further categorizes the range from 0.5 to 2 into three distinct segments, each corresponding to a different level of metacognitive prompt: Level 1 prompt is triggered for KMA index values from 0.5 to 1, Level 2 for values from 1 to 1.5, and Level 3 for values from 1.5 to 2, see Equation 5.2.

$$\text{Intervention's prompt} = \begin{cases} \text{No intervention}, & \text{if } 0 \leq \text{KMA index} < 0.5 \\ \text{Level 1}, & \text{if } 0.5 \leq \text{KMA index} < 1 \\ \text{Level 2}, & \text{if } 1 \leq \text{KMA index} < 1.5 \\ \text{Level 3}, & \text{if } 1.5 \leq \text{KMA index} \leq 2 \end{cases} \tag{5.2}$$

The Meta-Face Agent, using the M-FEI-based kernel, uses learners' facial cues to estimate MMP. This kernel sets thresholds for M-FEI's three outputs (see Section 4.7.4), $P(\texttt{precise})$, $P(\texttt{imprecise})$, and $P(\texttt{uncertain})$. Using the thresholds calculated in Section 4.7.4, the Agent delivers different levels of prompts. It is important to note that the prompts validated by Kautzmann & Jaques (2019) do not account for scenarios where learners are unable to articulate their confidence, i.e., `uncertain` MMP. Consequently, the M-FEI approach is employed to estimate the likelihood of `imprecise` MMP (i.e., the likelihood of being `imprecise` and the likelihood of not being `imprecise`). The Meta-Face Agent delivers metacognitive prompts across three levels based on the estimated $P(\texttt{imprecise})$: Level 1 prompt are triggered for scores ranging from 0.27 to 0.51, Level 2 for scores from 0.51 to 0.75, and Level 3

for scores from 0.75 to 1, see Equation 5.3.

$$\text{Intervention's prompt} = \begin{cases} \text{No intervention}, & \text{if } 0 \leq P(imprecise) < 0.27 \\ \text{Level 1}, & \text{if } 0.27 \leq P(imprecise) < 0.51 \\ \text{Level 2}, & \text{if } 0.51 \leq P(imprecise) < 0.75 \\ \text{Level 3}, & \text{if } 0.75 \leq P(imprecise) \leq 1 \end{cases} \tag{5.3}$$

5.3.4 Prompts in Metacognitive Intervention

In the Meta-Face Agent, the metacognitive intervention is strategically employed to encourage learners to reflect correctly on their performance. The tailored intervention is tiered into three distinct levels, designed to progressively deepen learners' metacognitive skills, as supported by past research (Kautzmann et al. 2016, Kautzmann & Jaques 2019, Wood & Wood 1999). Before the user study, the content of the prompts of the metacognitive intervention was developed in collaboration with a primary school mathematics teacher to ensure they are pedagogically sound and contextually appropriate for the target learners.

The first level of metacognitive prompts encourages learners to reflect abstractly on the question being asked, encouraging them to identify and consider the fundamental aspects of the query. This level aims to initiate the cognitive process that aligns with the onset of problem-solving, as shown in Figure 5.7b.

At the second level, the prompts are designed to help learners recall specific knowledge relevant to the question that has been previously taught in the classroom. This prompt facilitates the application of theoretical knowledge to practical problems, reinforcing the connection between classroom learning and real-world application, as illustrated in Figure 5.7c.

The third level of metacognitive prompts encourages learners to draw parallels between the current task and similar problems they have previously encountered or practiced in the classroom. This strategy, developed in collaboration with the primary school mathematics teacher, aims to consolidate knowledge and strengthen problem-solving skills by activating relevant prior experiences. An example of this prompt is shown in Figure 5.7d.

In scenarios where the Meta-Face Agent determines that metacognitive prompts

are not necessary, the system simplifies its interaction by reminding learners to proceed. In this case, the prompt is straightforward: 'Please submit your answer if you are ready.' This minimal intervention ensures that learners maintain control over their learning pace and decision-making process, providing autonomy in their educational journey.

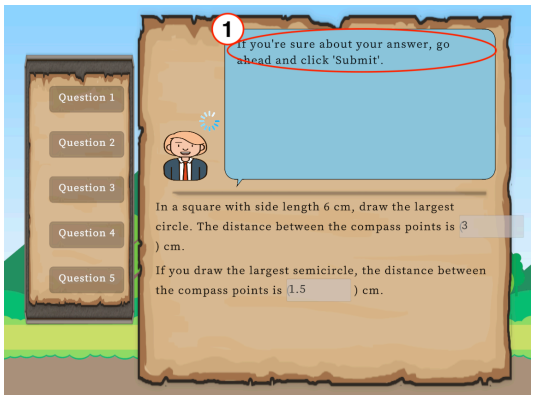
5.4 Evaluation study: Support Metacognitive Monitoring during Mathematical Exercises

This section outlines the recruitment procedure, provides an overview of the participants, and explains how groups were assigned in the user study. Ethical approval for this study was granted by the ethics committees of two universities (UK and China). This study recruited 215 pupils aged 7 to 12, an age range identified by Piaget's theory of cognitive development as the stage when children begin to perform logical operations mentally Piaget (1952). The study was conducted in collaboration with our research colleague, Dr. Kangcheng Wang, a psychologist specializing in child education in China.

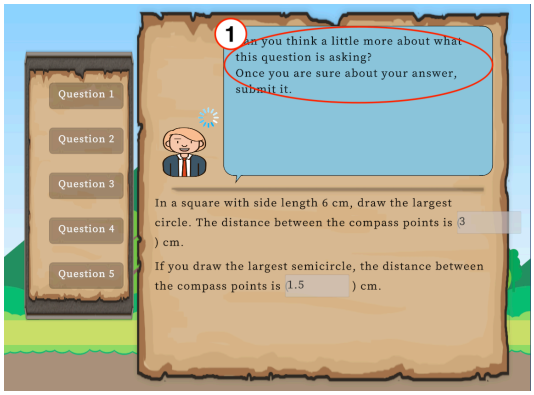
5.4.1 Selection and Participation of Children

The user study was conducted in December 2024 in China. We advertised this study in target schools. Prior to enrollment, a comprehensive overview of the study was provided, accompanied by tailored participant information sheets. It was clearly communicated that participation would have no impact on pupils' grades or school activities. The data management policy and participants' rights were thoroughly explained, with particular emphasis on the right to discontinue participation or withdraw data at any time without consequence.

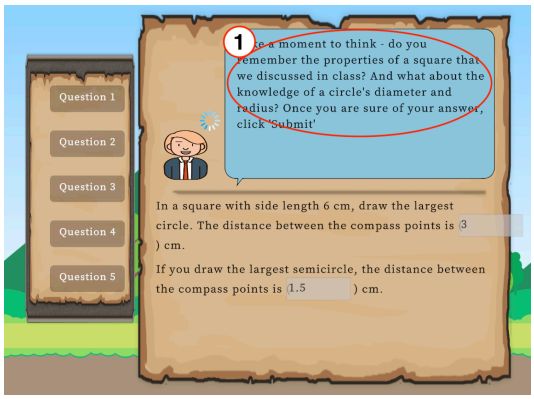
Participation in the user study was strictly voluntary. Pupils, along with their guardians, signed both the participant information sheet and the informed consent form. In the end, 235 pupils were initially invited to participate. Due to technical issues and pupil absences, data from 20 participants were deemed unusable. As a result, the final dataset comprised 215 pupils from Region A (109 females, 106 males; mean age = 9.51 years, SD = 1.48).



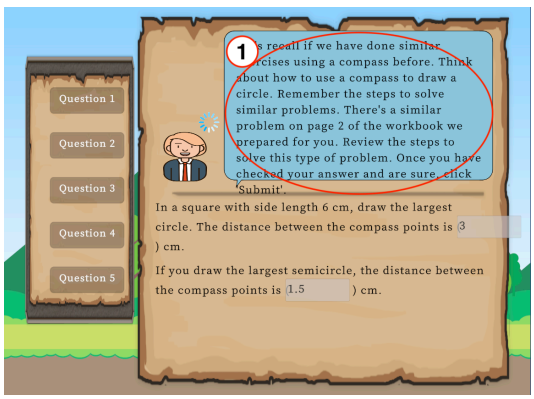
(a) No metacognitive intervention. Red circle: 'If you are sure about your answer, go ahead and click 'Submit'.'



(b) Metacognitive prompt Level 1. Red circle: 'Can you think a little more about what this question is asking? Once you are sure about your answer, submit it.'



(c) Metacognitive prompt Level 2. Red circle: 'Take a moment to think—do you remember the properties of a square that we discussed in class? And what about the knowledge of a circle's diameter and radius? Once you are sure of your answer, click 'Submit'.'



(d) Metacognitive prompt Level 3. Red circle: 'Let's recall if we have done similar exercises using a compass before. Think about how to use a compass to draw a circle. Remember the steps to solve similar problems. There's a similar problem on page 2 of the workbook we prepared for you. Review the steps to solve this type of problem. Once you have checked your answer and are sure, click 'Submit'.'

Figure 5.7: The metacognitive intervention in Meta-Face Agent

5.4.2 Materials

The Meta-Face Agent was configured (using the M-FEI or the KMA-based MMP estimation approach) for participants according to their assigned group (randomly allocated). All content was translated from English into Chinese to ensure accessibility in the pupils' native language. Sessions were conducted in classrooms at each participating primary school to provide a familiar and comfortable environment. Each classroom was equipped with desks, chairs, power chargers, laptops, and a whiteboard.

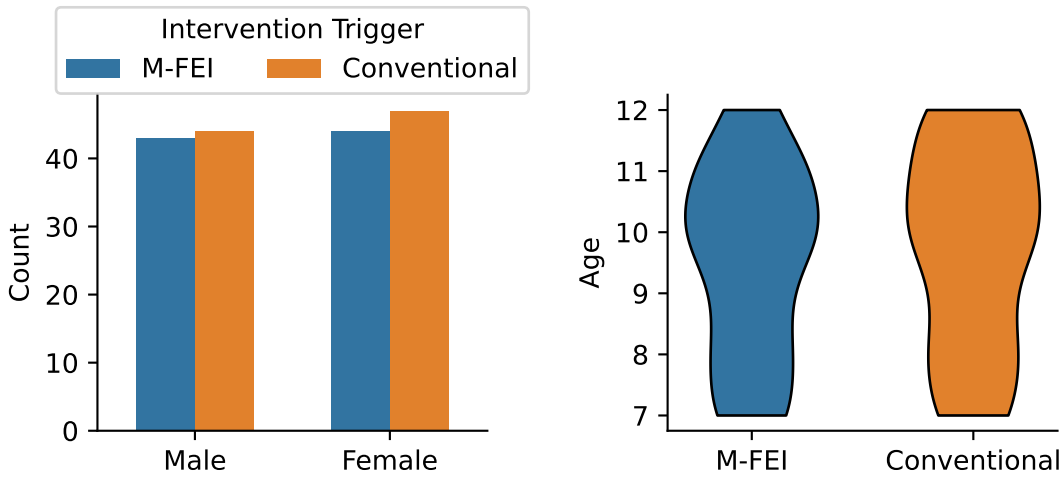
5.4.3 Procedure

In the user study, pupils engaged in a math-solving session within the Meta-Face Agent. We divided the participants into two groups. A total of 87 pupils were assigned to the 'M-FEI' group, where they completed the exercise with the metacognitive intervention using the 'M-FEI'. Conversely, 91 pupils were placed in the 'Conventional' group, completing the exercise with the intervention using the KMA-based approach. The math-solving session comprised five mathematical questions that pupils were required to solve in the order presented. When they input an answer to a question, the Meta-Face Agent prompted a message, 'How well do you think you answered?' Kautzmann et al. (2016), Kautzmann & Jaques (2019). Following this message, pupils had the autonomy to submit their answer directly or after revision.

As pupils were given two opportunities to answer each question, before and after the intervention, we conducted an additional test to investigate the learning effect associated with this repeated exposure factor (also called `practice` effect in research (Crawford et al. 1989, Song & Ward 2015)).

An additional test was conducted with participants from both the 'M-FEI' and 'Conventional' groups, of whom 72 pupils agreed to participate, 38 from the 'M-FEI' group and 34 from the 'Conventional' group. In addition, 37 new pupils were recruited to form a 'Control' group. During this phase, all participants were asked to solve five new, but similar, mathematical questions without receiving any intervention.

For clarity, the terms `intervention` factor and `practice` factor are used in the following content to distinguish the conditions being examined.



(a) Gender distribution

(b) Age distribution

Figure 5.8: The demographic distributions in the 'M-FEI' group and the 'Conventional' group. These groups' experience and interactions were shaped by the M-FEI and the KMA-based MMP estimation approach.

5.4.4 Data Collection

Both video recordings and system logs generated by the Meta-Face Agent were collected for analysis. All collected data are securely stored on a password-protected, encrypted server for a maximum duration of two years. Access to the data is restricted to the authors of this study.

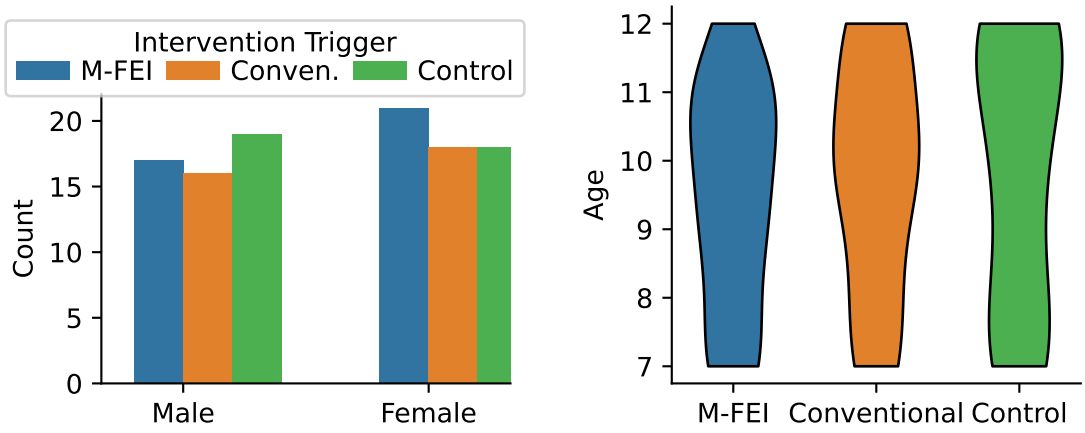
5.5 Data Preparation

This section presents the demographic information of the user study participants by group and describes the measurements used for key variables in the study.

5.5.1 Demographic Information in Groups

In the user evaluation study, participants were randomly assigned to either the 'M-FEI' group or the 'Conventional' group. The demographic distributions of these two groups are presented in Figure 5.8. Both groups contain nearly equal numbers of male and female participants. Regarding the age distribution, both groups have more participants aged around 11, with fewer at the youngest and median ages.

An additional test was conducted to analyze pupils' learning outcomes when they



(a) Gender distribution

(b) Age distribution

Figure 5.9: The demographic distributions in the additional test. Note: No intervention is given in the three groups of the additional test. The 'M-FEI' and the 'Conventional' groups have pupils who completed exercises with the intervention tailored by M-FEI-kernel and KMA-kernel, respectively. The 'Control' group has pupils who complete exercises for the first time.

encountered similar maths questions for a second time. The demographic distribution of the groups involved in this post-test is shown in Figure 5.9.

5.5.2 Variable Notations and Measurements

Exercise score ES: Five maths questions were designed for the math-solving exercise in the user evaluation study. All math questions were selected from the math-question bank, which matched with the teaching materials learned by our participants in the 'Shandong Education Edition Mathematical Textbook' curriculum *Shandong Education Press* (2025). The selections consist of two easy (1 mark), one medium-level (2 marks), and two hard-level (3 marks) questions. Prior to the user study, all (i.e., four) mathematics teachers (grades 1 to 6) of the participating pupils were invited to rate the difficulty of the selected questions (1=easy, 5=hard) and to give their assessment of how well these aligned with the teaching curriculum (1=unrelated, 5=highly related). The questions received an average difficulty rating of 3.25 and an average curriculum relevance rating of 4.5.

The exercise scores were calculated using the formula presented in Equation 5.4. Pupils' scores before and after the intervention are denoted as ES_{stand} and ES_{rev},

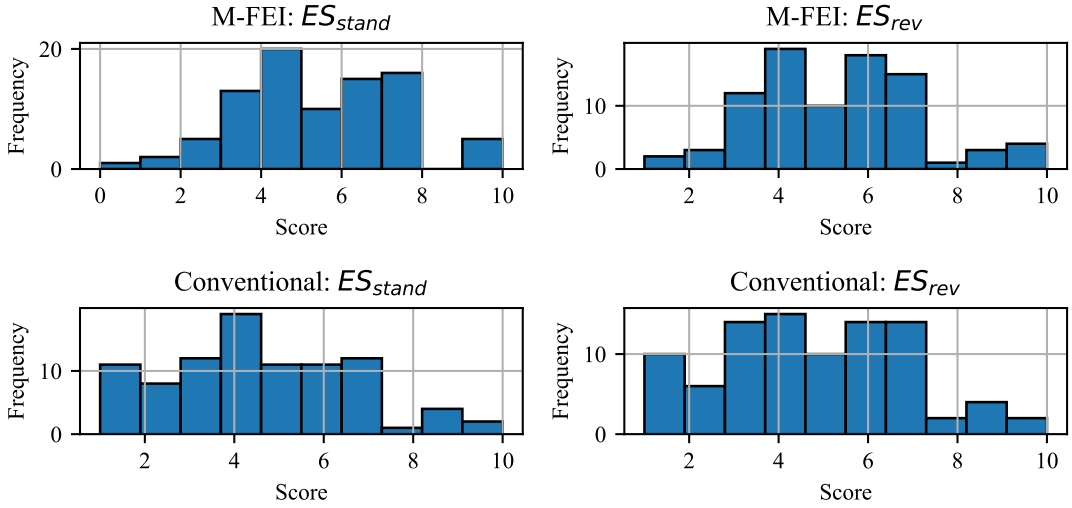


Figure 5.10: The distribution of pupils' exercise scores in Meta-Face Agent.

respectively.

$$I_i = \begin{cases} 0, & \text{if question } i\text{'s answer is wrong,} \\ 1, & \text{if question } i\text{'s answer is correct.} \end{cases} \tag{5.4}$$

$$PS = 1 \times I_1 + 1 \times I_2 + 2 \times I_3 + 3 \times I_4 + 3 \times I_5 \tag{5.5}$$

5.6 Results

Pupils' learning outcomes (ES_{stand} and ES_{rev}) were examined in relation to the tailored intervention using M-FEI† and CA-4[3]. The distributions of dependent variables, ES_{stand} and ES_{rev}, are illustrated in Figure 5.10.

Given the limitations of the validated intervention by Kautzmann et al. for `imprecise` MMP estimation, the M-FEI† was configured to identify `imprecise` MMPs. The intervention strategies from Section 5.3.3 are illustrated in Figure 5.11.

Prior to analyzing learning outcomes, demographic variables were examined across groups. No significant difference in gender distribution was observed between the 'M-FEI' and 'Conventional' groups, as indicated by a chi-square test ($\chi^2 = 0.0$, df $= 1$, p

[3]M-FEI without AHG, as proposed in Ruan et al. (2025) and denoted M-FEI†, was selected for the user evaluation study due to its lower system workload (eliminating the need to process facial cues via OpenFace). It has comparable performance to M-FEI* among Region A pupils. Given that the math-solving exercise consisted of five questions, CA-4 was adopted as the conventional approach, following Kautzmann et al. (2016), Kautzmann & Jaques (2019).

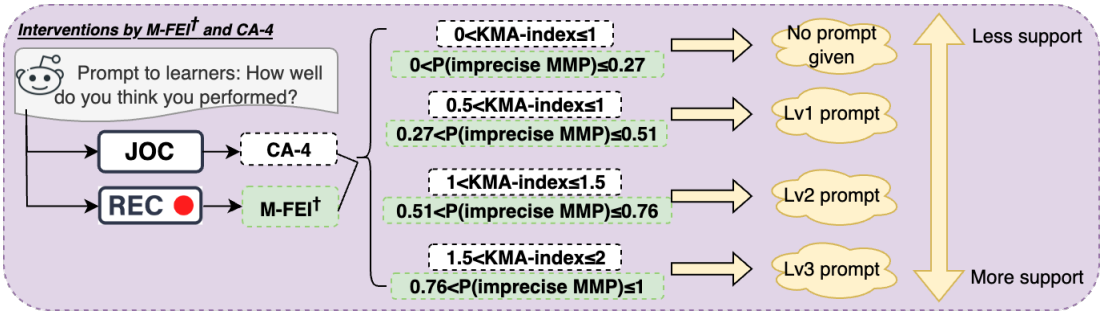


Figure 5.11: The strategy of launching the metacognitive prompt

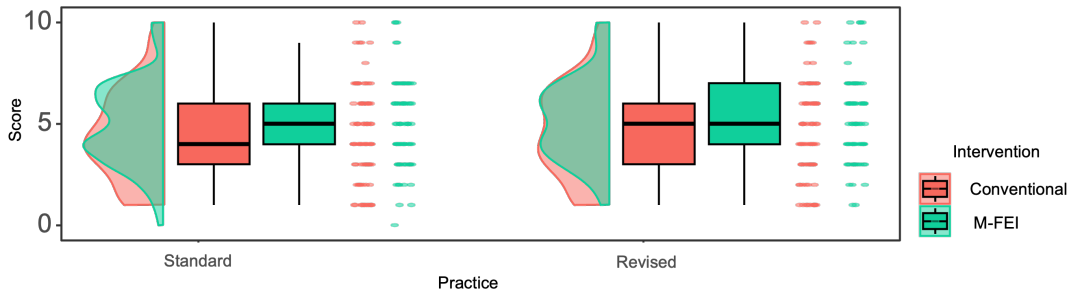


Figure 5.12: Two-way ANOVA: Effects of the `practice` factor and the `intervention` factor on exercise score (ES).

= 0.99) and a t-test (t(176) = –0.14, p = 0.89). Similarly, age distributions did not differ significantly between the groups, as shown by the Kolmogorov–Smirnov test (D = 0.08, p = 0.88) and a t-test (t(176) = –0.86, p = 0.39).

A comparison of ES_{stand} between the 'M-FEI' and 'Conventional' groups using a t-test revealed no significant difference (t(176) = 1.45, p = 0.15). To further examine the effects of the `intervention` and `practice` factors, illustrated in Section 5.4.3, on math-solving outcomes, a two-way ANOVA was conducted (see Figure 5.12)(Field 2013). The results indicated a significant main effect of the `intervention` factor (F(1, 176) = 5.00, p = 0.03). In contrast, the `practice` factor did not show a significant effect (F(1, 176) = 1.31, p = 0.25), nor was there a significant interaction between the two factors (F(1, 176) = 0.04, p = 0.85).

To address RQ2.1, a paired t-test was conducted to compare exercise scores between ES_{stand} and ES_{rev} for both the 'M-FEI' and 'Conventional' groups. The results indicated that ES_{rev} was significantly higher than ES_{stand} in both groups. In the 'M-FEI' group, a significant increase in the average exercise score was observed (t(86) = 3.45, p < .01). Similarly, the 'Conventional' group also showed a significant improvement (t(90) = 2.13, p = .02). The reports for both paired t-test is illustrated

Table 5.1: Comparison of ES_{rev} and ES_{stand} in M-FEI and Conventional Groups. In both groups, ES_{rev} was significantly higher than ES_{stand}, indicating improved performance when learners received tailored interventions using both approaches.

Group	Measure	Mean	t-value	p-value	Result
M-FEI	ES_{stand}	4.95	3.45 (df = 86)	$< .01$	$ES_{rev} > ES_{stand}$
	ES_{rev}	5.26			
Conventional	ES_{stand}	4.48	2.13 (df = 90)	.02	$ES_{rev} > ES_{stand}$
	ES_{rev}	4.70			

in Table 5.1.

Furthermore, to address RQ2.2, an independent t-test comparing ES_{rev} across groups revealed that the 'M-FEI' group outperformed the 'Conventional' group, with a statistically significant difference ($t(176) = 1.71$, $p = .04$), see Table 5.2.

Table 5.2: Comparison of ES_{rev} between M-FEI and Conventional Groups. A significant difference was observed, with the M-FEI group showing higher average performance.

Group	Measure	Mean	t-value	p-value	Result
M-FEI	ES_{rev}	5.26	1.71 (df = 176)	.04	'M-FEI' sig. higher
Conventional	ES_{rev}	4.70			

In the additional test, 38 pupils from the 'M-FEI' group and 34 from the 'Conventional' group, along with 37 pupils from a control group, participated. Their exercise scores are presented in Figure 5.13. An ANOVA test revealed no significant differences in exercise scores among these three groups ($F(2, 106) = 0.13$, $p = 0.88$), see Table 5.3.

5.7 Discussion

Building on the results presented in Section 5.6, this section discusses the educational benefits of tailoring interventions using the M-FEI-based MMP estimation approach. It then extends the discussion to consider key intervention factors that may support

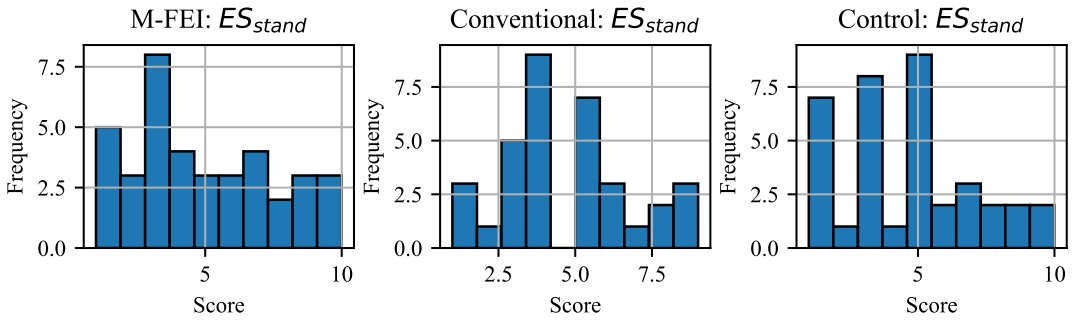


Figure 5.13: The distribution of pupils' exercise scores in Meta-Face Agent (additional test).

Table 5.3: ANOVA test comparing mathematical outcomes across M-FEI, Conventional, and Control groups. No significant effect of the practice factor was found.

Group	Measure	F-value	p-value	Result
M-FEI	ES_{stand}	0.13 (df = 2, 106)	.88	`Practice` factor has no effect
Conventional	ES_{stand}			
Control	ES_{stand}			

the development of metacognitive skills over the long-term study.

5.7.1 Educational Benefits of Launching the Metacognitive Intervention using M-FEI

Unlike KMA-based approaches that rely on learners explicitly articulating their confidence (e.g., rating their answers as correct or incorrect), the proposed approach, M-FEI, eliminates the need for learners to complete the JOC questionnaire during the metacognitive intervention. Instead, it encourages learners to engage in self-reflection on their performance by answering an open-ended prompt ('How well do you think you answered?'), which supports metacognitive engagement and maintains learning flow without requiring forced-choice confidence ratings (Riku 2021). This addresses a key challenge faced by young learners, who often struggle to articulate their confidence levels (Harter 2012, Harris & Brown 2013, Amershi et al. 2014, Lehnert 2024).

The results for RQ2.1 underscore the educational benefits of tailoring the targeted intervention based on both the M-FEI and KMA-based approaches. Notably, tailoring

the intervention using the M-FEI method led to more substantial improvements in learning outcomes, i.e., smaller p value in the t-test. This effect is particularly evident in the analysis for RQ2.2, the comparison of standard and revised exercise scores (ES_{stand} and ES_{rev}). Results showed that the M-FEI group achieved a significantly higher ES_{rev}, indicating enhanced learner performance following the intervention.

Furthermore, the analyses from our study indicate that the positive influence on mathematical learning outcomes can predominantly be attributed to the effects of the intervention rather than to practice effects. This finding is significant as it aligns well with existing literature that distinguishes genuine learning gains from mere practice effects (Duff et al. 2010). These insights substantiate the efficacy of the M-FEI approach in facilitating learners' MMP and then boosting their mathematical learning outcomes.

In light of these results, broader implementation of the M-FEI method is recommended in educational settings where MMP is critical. Although the present study was conducted within the context of mathematical learning, MMP is recognized as a fundamental component across a range of educational domains. Accordingly, future research should investigate the applicability and effectiveness of the M-FEI method in other learning contexts, such as science education, language learning, and social studies, to assess its generalizability and its impact on learning outcomes in disciplines where metacognitive strategies play a central role.

The math-solving exercise in this study consisted of five questions, limiting each pupil to a maximum of five rounds of intervention interactions. This relatively brief exposure constrains the ability to fully assess the potential educational benefits of the interventions. Prior research has emphasized the need for sustained and frequent interventions to produce meaningful improvements in metacognitive skills (Dignath & Büttner 2008). Therefore, future studies should consider implementing longer-term interventions using M-FEI to more effectively evaluate its impact on the development of metacognitive skills. Such investigations may offer deeper insights into the optimal frequency and duration of interventions required to achieve measurable and long-lasting educational outcomes.

5.7.2 Impact of Intervention Duration and Frequency on Learning Skills

How long and how often the intervention should be delivered to learners (its duration and frequency) so that influences its effectiveness in developing learning skills. A long-term goal of our research is to enhance learners' metacognitive monitoring skills, i.e., improve their MMP even without metacognitive interventions.

In pedagogical theory, providing students more practice opportunities (for example, encouraging them to do metacognitive monitoring based on the given prompts in Section 5.3.1) should reinforce skills. Some prior research concluded that interventions with more sessions in total tend to achieve larger learning effects (Dignath & Büttner 2018). This is aligned with cognitive principles, the spaced repetition (Ebbinghaus 2013), distributing more intervention practices over time yields better learning gains. However, recent studies also claimed that more practices are not beneficial indefinitely. The review research done by Wang & Sperling (2020) reports that there is no clear relationship between total intervention practices and learners' outcomes. It suggests that the relationship between the duration of engagement with intervention practices and learning outcomes is mixed. It is clear that extremely prolonged interventions can suffer from waning student interest or plateauing benefits. Thus, future work to enhance metacognitive skills should be aware of the duration of the total intervention.

Regarding the frequency of intervention sets, the evidence is also mixed. Sessions with very high frequency (for example, daily sessions) can accelerate initial learning, but if too intensive, they risk cognitive overload or boredom. Conversely, very low sessions (monthly) may not provide enough opportunities to develop skills. In the meta-analysis of Eberhart et al. (2025), there is no significant correlation between the frequency of sessions (how many sessions per week) and interventions' outcomes. This implies that, within a typical duration, how many of sessions per week delivered did not systematically change the effectiveness. Thus, the extremely low or high frequencies are generally avoided. Consider metacognitive skill development, which involves reflection and gradual habit-building, allowing time between sessions for students to assimilate and apply strategies in class, might actually bolster effectiveness (Benjamin & Tullis 2010). The frequency of giving practices for enhancing metacognitive monitoring should be carefully considered.

5.8 Limitations and Future Works

A notable limitation of this study lies in the reliance on the existing validated metacognitive intervention, which was specifically designed for addressing `imprecise` MMP. This constraint limited the potential of the M-FEI approach, which is capable of distinguishing between `precise`, `imprecise`, and `uncertain` MMP categories. As a result, the current user evaluation did not explore or implement advanced intervention strategies that fully leverage M-FEI's broader capabilities.

Future work should expand the scope of intervention design by developing a broader set of metacognitive strategies that can be selected or tailored more precisely using the feedback adapted by M-FEI. While M-FEI does not author interventions, it can support expert educators or ITS designers in selecting and deploying appropriate interventions by providing detailed, real-time indicators of learners' metacognitive states. Additionally, these insights can be used to trigger large language model (LLM)-based solutions to automatically generate contextually relevant, text-based prompts or guidance, further enriching the learning experience with personalized metacognitive support. Such efforts would enhance intervention effectiveness across diverse learning contexts and further validate the utility of M-FEI in dynamically responding to learners' varying metacognitive needs. Additionally, systematic evaluation of these new interventions in varied educational settings is essential to establish their broader potential impact on learning outcomes.

In addition, this PhD study included an inter-regional evaluation in Chapter 4 to validate the performance of M-FEI among Scottish pupils. Promising results were observed within this population. Therefore, future work should focus on implementing the Meta-Face Agent in the context of mathematics learning aligned with the Scottish curriculum. In this implementation, metacognitive interventions should be tailored to support Scottish pupils during math problem solving, informed by M-FEI's real-time analysis of metacognitive states. Based on the demonstrated effectiveness of M-FEI in this study, it is anticipated that similar improvements in mathematical learning outcomes could be achieved in broader applications within the Scottish educational setting.

5.9 Chapter Summary

This reported user evaluation study reveals the feasibility of interpreting facial expressions to estimate MMP in children. The Meta-Face Agent was implemented to tailor the metacognitive intervention based on both the M-FEI and KMA approaches.

Findings indicate that the intervention tailored by the M-FEI approach results in more significant improvements in math-solving outcomes. In addition, the potential influence of the practice effect was examined, and the results confirmed that the observed gains were attributable to the metacognitive interventions rather than repeated exposure alone.

The study also highlights the importance of considering the duration and frequency of metacognitive interventions in the development of metacognitive skills. To this end, a long-term training approach is recommended to more effectively foster these skills, with careful attention to the scheduling and intensity of practice sessions.

To be clear, this work makes the following three unique contributions:

1. Implements and evaluates the Meta-Face Agent within a young learner population, showcasing the effectiveness of tailoring metacognitive interventions based on both the M-FEI and KMA approaches.

2. Establishes that metacognitive interventions guided by the M-FEI approach lead to greater improvements in mathematics problem-solving performance among children aged 7 to 11, compared to interventions based on self-reported measures.

3. Confirms that the observed performance gains in young pupils are primarily attributable to metacognitive interventions rather than practice effects, strengthening the causal link between intervention and learning outcomes.

Chapter 6

Conclusion

Hope is like the sun, which, as we journey toward it, casts the shadow of our burden behind us.

— Samuel Smiles

This chapter is the conclusion of this thesis. It begins by revisiting the two main research questions and summarizing the key findings related to each sub-question. It then outlines the primary contributions of this PhD research, followed by a discussion of the identified limitations. Following that, the chapter presents directions for future work and opens questions for further investigation. In the end, it concludes with a Looking Ahead section, which reflects on the broader educational implications of this research and envisions how AI-driven metacognitive support systems may shape future learning environments.

6.1 Key Findings of Research Questions

This thesis investigates two main questions, RQ1 and RQ2, and their corresponding sub-questions are summarized below, followed by a discussion of the findings for each question in the subsequent subsections.

6.1.1 RQ1: To what extent does the metacognitive monitoring performance (MMP) impact learning outcomes, and can pupils' MMPs be inferred from their facial cues?

6.1.1.1 RQ1.1: How does MMP influence pupils' task scores?

This question is directly addressed in Chapter 3. The correlation analysis conducted in Section 3.8.1 shows that pupils' MMP positively influences their performance on maths-related cognitive tasks. These findings extend prior research by revealing this effect specifically in young learners aged 7 to 11 within computer-based learning environments (CBLEs). Pupils who more accurately monitored their confidence levels on their problem-solving processes achieved higher task scores.

6.1.1.2 RQ1.2: What facial cues have a significant correlation with MMP?

This question was addressed in Section 3.8.2. The results identify a range of facial cues that significantly correlate with MMP, as presented in Table 3.5 and Table 3.6. Key correlated cues include inner brow raiser, brow lowering, lips parting, cheek raising, lip corner pulling, gaze directions (up-down, left-right), head nodding, and head movements (up-down, left-right).

Moreover, gender acts as a factor influencing these correlations. For male pupils, significant indicators include brow lowering, eyelid tightening, lip tightening, lips parting, lip corner depressor, upward gaze, and horizontal head movements. In contrast, for female pupils, horizontal and vertical gaze shifts, as well as vertical head movements, show significant correlations with MMP.

These findings are the first contribution to the growing understanding of facial cues associated with MMP in pupils aged 7 to 11 during math-related tasks in CBLEs. They provide valuable insights for estimating MMP through non-verbal expressions.

6.1.1.3 RQ1.3: What is the performance of the conventional approach in estimating pupils' MMP?

To answer this question, Section 4.7.2 presents an evaluation of a conventional method for estimating MMP using the Knowledge Monitoring Assessment (KMA) matrix (see Section 4.6.1). The results are the first empirical evidence revealing the

performance of the conventional approach. It shows that the approach's accuracy improves as the number of samples in the KMA matrix increases, but eventually reaches a plateau, indicating a saturation point in predictive performance. Additionally, the method's results reveal consistent accuracy across different populations, specifically pupils from two Chinese provinces and those from Scotland, highlighting its generalizability across diverse cultural contexts.

6.1.1.4 RQ1.4: Is it possible to estimate pupils' MMP using deep learning to interpret their facial expressions? If yes, does this improve MMP estimation?

Section 4.7.5 reports 'Yes' to this question. It provides the first empirical support for the feasibility of using deep learning to estimate MMP from facial expressions. The developed Meta-Facial Expression Interpreter (M-FEI) showed superior performance compared to random guessing, validating its capability to identify MMP states. In comparison to the conventional KMA-based method, M-FEI achieved improved estimation accuracy and significantly reduced false alarms, particularly in detecting `imprecise` MMP. Although its effectiveness decreased somewhat in inter-regional validation, the model still outperformed the conventional approach in identifying MMP and maintained a consistently low false alarm rate. These findings confirm both the viability of the deep learning approach and its advantages over the conventional approach in enhancing MMP estimation.

6.1.1.5 Summary for RQ1

The findings confirm a significant positive correlation between young pupils' MMP and their maths-related learning outcomes in CBLEs. However, the relationship between facial cues and MMP is highly complex, making linear predictive models inadequate. To address this, the deep learning-based M-FEI model was developed and shown to effectively estimate pupils' MMP by interpreting their facial expressions. Moreover, M-FEI achieves higher estimation accuracy and produces fewer false alarms compared to conventional approaches, maintaining its effectiveness even across different cultures, as confirmed by validations in Chapter 4.

6.1.2 RQ2: Given the established benefits of metacognitive interventions in educational research, can interventions tailored to MMP, as identified through facial cues, enhance pupils' mathematical learning outcomes in CBLEs?

6.1.2.1 RQ2.1: Does the intervention using the M-FEI approach improve pupils' mathematical learning outcomes?

This question is investigated in Section 5.6, which presents the outcomes of a mathematical exercise administered before and after intervention. The results show a statistically significant improvement in pupils' average scores following the metacognitive monitoring interventions tailored by M-FEI. This finding suggests that the M-FEI not only assists in accurately identifying MMP but also facilitates the adaptation of timely and individualized metacognitive support, thereby enhancing their mathematical learning outcomes.

6.1.2.2 RQ2.2: How do the mathematical learning outcomes of pupils who receive the tailored intervention using the M-FEI approach compare with those of pupils who undergo the KMA-based approach?

This comparison is examined in Section 5.6, where two pupil groups matched for gender, age, and mathematical performance were evaluated. The results reveal that pupils who received tailored metacognitive monitoring interventions via the M-FEI approach achieved significantly higher average scores than those supported by the conventional KMA-based method. This suggests that the M-FEI approach not only enhances young pupils' MMP but also leads to more effective instructional support, resulting in superior learning outcomes.

6.1.2.3 Summary for RQ2

The results support that metacognitive interventions tailored to pupils' MMP, as identified through facial cues using the M-FEI approach, can enhance mathematical learning outcomes. Chapter 5 shows that pupils receiving M-FEI-guided interventions experienced significant improvements in post-intervention mathematical performance, illustrating the effectiveness of timely, individualized metacognitive support.

Additionally, when compared to the conventional KMA-based approach, the M-FEI method resulted in higher learning gains among pupils, confirming its advantage in both accurately identifying MMP and delivering more targeted instructional support.

6.2 Key Contributions

This research involved two user studies followed by analyses of data from those studies. This section will highlight key contributions of this project, including understanding MMP, M-FEI, and practical implementations, with limitations discussed in the following section.

6.2.1 Understanding MMP through Facial Cues

Metacognitive monitoring, often described as 'thinking about thinking,' refers to the internal process of evaluating one's own understanding and performance. It often operates behind the scenes, subtle and difficult to detect directly. While prior research has attempted to link emotional experiences to metacognitive processes, identifying correlations between certain emotions and MMP, this thesis advances that understanding by focusing on facial cues.

This thesis shows that learners' facial expressions, gaze direction, and head movements are not only correlated with MMP but can also serve as predictive signals. These diverse facial cues collectively indicate that MMP is reflected by multimodal responses. Furthermore, the relationship between facial expressions and MMP is influenced by two key factors: gender and task-specific cognitive capacity. Experimental results presented in Section 3.10.3 illustrate the complexity of this relationship and highlight the potential advantage of employing deep learning techniques to effectively estimate MMP from facial cues.

6.2.2 M-FEI

M-FEI represents the core contribution of this thesis. It is the first deep neural network model to identify MMP through facial expressions. It was developed to estimate `precise`, `uncertain`, and `imprecise` MMP and is optimized for local deployment on low-performance devices, such as CPU-based laptops, enhancing its practical applicability while preserving learners' privacy. By enabling all computations to run

locally on the learner's own equipment, the approach releases the need to transmit sensitive facial data to external computational servers, ensuring privacy protection without compromising performance.

Extensive numerical experiments were conducted to evaluate the effectiveness of M-FEI. The results show that M-FEI significantly improves the identification of MMP, achieving higher accuracy and a lower false alarm rate compared to the conventional approach. In inter-regional validation experiments, M-FEI was first trained on data collected from one Chinese province and then tested on data from a different Chinese province and from Scotland. Even under these challenging conditions, M-FEI consistently outperformed the conventional approach, maintaining higher accuracy and lower false alarm rates.

This contribution also offers insight for future research: facial cues indicative of MMP may exhibit shared patterns across different cultural contexts, suggesting a level of cross-cultural generalizability.

Furthermore, the second user study in this research revealed the educational benefits of M-FEI. Pupils achieved higher mathematical learning outcomes when guided by interventions tailored through M-FEI, highlighting its potential to inform adaptive learning strategies.

Given its illustrated advantages, M-FEI was exported to the ONNX format (ONNX Community 2025), enabling platform-independent deployment. A C-sharp script was developed to execute the model, and the implementation has been made publicly available on GitHub (github/affect2mmp 2025). Researchers can utilize M-FEI to identify MMP via standard input/output interfaces. Importantly, M-FEI was trained on data collected from pupils performing cognitive skill tasks, making it a flexible framework suitable for deployment across various CBLEs, such as smart games, intelligent tutoring systems (ITSs), and other interactive learning platforms. Its generalizable design supports use in a wide range of subject areas, including mathematics, science, and language learning, reflecting the domain-independent nature of metacognitive monitoring skills.

6.2.3 Practical Implementations

6.2.3.1 Meta-Brainhood

Meta-Brainhood was developed to encourage pupils to engage in metacognitive monitoring while simultaneously collecting facial recordings and log data related to their MMP. To support broader research efforts, Meta-Brainhood is available in both Chinese and English versions. This prototype can be used by future researchers to gather facial data for further exploration of MMP.

6.2.3.2 Meta-Face Agent

The Meta-Face Agent developed in this research is an ITS designed to deliver real-time, low-disruption metacognitive monitoring interventions to learners[1]. It integrates M-FEI to dynamically tailor support based on learners' ongoing MMP, allowing interventions to be timely and contextually relevant without interrupting their learning flow.

This prototype contributes a valuable advancement to HCI, particularly in the domain of Child-Computer Interaction (CCI). Children often struggle to articulate their internal cognitive states or emotions, making it difficult for traditional systems that rely on verbal self-reports to adapt effectively. By leveraging facial cues, the Meta-Face Agent bypasses this limitation, allowing the system to infer metacognitive states non-verbally and respond with personalized support.

The Meta-Face Agent suggests new possibilities for both educators and researchers, moving beyond self-reported responses toward systems that can interpret learners' cognitive states implicitly and adjust support accordingly.

6.3 Limitations

This section outlines five key limitations of the research: the general limitation of error in estimation, challenges in data labeling, the narrow focus on specific aspects of MMP, and limitations related to both generalizability and subject-specific scope.

[1]Work related to the Meta-Face Agent is currently under peer review. The system and its associated findings will be made publicly available upon acceptance of the corresponding publication.

6.3.1 Limitation of Data Labeling and Measurement Validity

A primary limitation of this PhD research, as discussed in Chapter 3, lies in the reliance on self-reported confidence ratings for data labeling. Although commonly used, self-reported measures can be unreliable due to factors such as social desirability bias. In such cases, participants may provide responses they believe are more acceptable or favorable rather than accurately reflecting their internal states. This limitation may introduce inaccuracies in the data labels, which can affect the performance of the M-FEI method in estimating pupils' MMP. To address this concern, future studies are encouraged to refine data collection procedures, invite expert teachers, and consider the inclusion of more objective indicators of learner confidence, such as physiological signals (e.g., heart rate variability) (Liu et al. 2008).

6.3.2 Limitation of Accuracy of Proposed M-FEI

The limitation of accuracy is a broader concern common to many detection systems, including the one proposed in this PhD research. While the M-FEI technique outperforms the conventional KMA-based approach in terms of accuracy and reduced false alarm rates, its estimations remain subject to occasional misclassifications. Even small inaccuracies can lead to meaningful consequences in educational contexts. For instance, in the estimation of MMP, there is a risk of falsely identifying learners with precise MMP as requiring support, or failing to detect learners with imprecise MMP who would benefit from intervention. To mitigate such risks, careful consideration is needed in how M-FEI is applied in future systems. Rather than relying on fully automated decisions, the system could be designed to supplement teacher judgment or trigger low-stakes scaffolds, such as providing optional reflective prompts or offering gentle strategy reminders that do not directly affect learners' performance evaluations. As discussed in Chapter 4, strategies such as setting appropriate decision thresholds, incorporating confidence scores, or implementing human-in-the-loop mechanisms may help reduce the impact of occasional misclassifications and promote more responsible deployment.

6.3.3 Limitation of Narrow Temporal Focus in Metacognitive Monitoring

The study focused on examining pupils' MMP during reflections on completed tasks. However, metacognitive monitoring also occurs prior to task engagement, when learners forecast their expected performance. This limitation indicates that the anticipatory dimension of metacognitive evaluation was not captured, despite its potential influence on learners' strategic planning and academic outcomes. To achieve a more comprehensive understanding of MMP, future research should incorporate both retrospective and prospective aspects of metacognitive monitoring.

6.3.4 Limitation of Cultural Diversity

Although results from inter-regional validation suggest that the M-FEI method may be applicable to both China and Scotland regions, the generalizability of these findings remains limited. Variations in facial expressions associated with metacognitive monitoring across different ethnic and cultural backgrounds may affect the accuracy of MMP estimation, thereby restricting the applicability of the results to broader demographic groups. To address this limitation, future research should incorporate more diverse datasets, including participants from varying age groups, ethnicities, and other demographic, economic, and social factors, to ensure that the M-FEI method can effectively estimate MMP across a wider population.

6.3.5 Limitation of Context Constraints in User Evaluation

The user study presented in Chapter 5 was conducted within the context of mathematics education, where tailored metacognitive monitoring interventions were implemented using both the M-FEI method and the conventional approach. Due to resource constraints, the study was not extended to other subject areas. However, given that metacognitive monitoring plays a vital role across various domains of learning, this represents a limitation in terms of the study's broader applicability. Future research should address this limitation by extending the implementation and evaluation of MMP-based interventions to other subject areas, such as science, language learning, and the subjects related to the humanities, to examine the generalizability and effectiveness of these approaches in diverse educational contexts.

6.3.6 Limitation of Intervention Duration and Frequency in User Evaluation

Another limitation of the user evaluation study lies in the short duration and single session of learners' interactions with the Meta-Face Agent. In typical educational settings, learners engage with ITS platforms over extended periods and return repeatedly to the learning environment, often revisiting topics and tasks at varying frequencies to reinforce understanding through repetition. This iterative process of spanning days, months, or even years is critical for supporting long-term learning gains and skill development. The limited intervention in this study does not fully capture these longer-term usage patterns, which likely influence both learners' learning gains and the instructional adaptations made by educators. Such long-term, repeated interactions also introduce development requests to human-computer interaction (HCI) research, necessitating more adaptive and sustained support mechanisms than the current study.

6.3.7 Ethical Limitation

The use of systems implemented with M-FEI in schools raises important ethical questions. When applied responsibly, metacognitive monitoring can provide teachers with insights that help them offer timely, individualized support, particularly in large classrooms where personal attention is limited. Used in this way, the technology empowers pupils by fostering self-awareness and strengthening their ability to reflect on their own learning.

Even though M-FEI is a fully local deployment model that processes data on the device without transferring it to remote servers, ethical risks remain. Continuous monitoring of facial expressions, gaze, and gestures could still create a sense of excessive surveillance or pressure pupils to conform to expected patterns of behavior. To remain ethical, such systems must be used with transparency, clear educational purposes, and safeguards that ensure they support rather than control learners.

6.4 Future Work

This section brings together some future work and questions to further investigate.

6.4.1 Establish Platform for Exchanging MMP Data

Previous research, including the work of Jack et al. (2012), has demonstrated that facial cues associated with affective states can differ markedly across cultural contexts. These cultural variations highlight the risk of misinterpreting affective states when facial interpretation techniques are applied without appropriate cultural calibration. This opens a new question based on this work: How do cultural norms in emotional expression affect the interpretation of metacognitive facial cues?

To enhance the accuracy and applicability of MMP estimation across diverse cultural contexts, future research should focus on the creation of a collaborative platform. This platform would serve as a repository and exchange hub for researchers to share, compare, and standardize statistical metrics and measurements of facial cues and other physiological signals (such as average and variance of action units, gaze directions, and head gestures) that relate to metacognitive monitoring. Such a resource would not only facilitate cross-cultural validation of FER techniques but also encourage replication studies and help identify culturally specific patterns in physiological data.

Whilst maintaining the critical design constraint of safeguarding learners' personal data, it remains feasible to capture and share non-identifiable behavioral indicators through commonly available local devices. In this context, audio signals represent a valuable yet underexplored modality that should eventually be investigated and incorporated into the data-sharing framework. These audio data could either be stored alongside video recordings within the same repository or shared independently in cases where video data is subject to stricter privacy restrictions. This extension would facilitate more comprehensive multi-modal analyses while respecting privacy considerations.

Ultimately, these efforts would significantly enhance the generalizability and robustness of MMP estimation based on human behavior data, strengthening its relevance in both educational and psychological research. Ensuring cultural sensitivity in these methodologies is essential for improving their reliability and effectiveness in real-world applications.

6.4.2 Human's Preference on Judgments of Confidence

Since self-labeled data can be susceptible to label noise, future research should prioritize enhancing the quality and reliability of the Affect2Metacognition dataset through improved labeling procedures.

To address this issue, future efforts should consider involving expert tutors or teachers in the evaluation of pupils' JOC responses and their observed behaviors. Expert involvement could substantially reduce the risk of label noise by offering more objective assessments of pupils' metacognitive states. In cases where experts disagree, consensus discussions or averaging their assessments could help improve labeling reliability. Additionally, the implementation of a standardized rating framework would further enhance the robustness of the dataset, increasing its value as a resource for advancing the understanding and improvement of metacognitive processes in educational contexts.

Furthermore, exploring automated or semi-automated systems that assist in labeling while incorporating teacher insights could offer a scalable solution to balance objectivity with resource efficiency. Such systems could potentially leverage advanced analytics to pre-filter or suggest labels, which could then be refined or confirmed by human experts.

6.4.3 New Metacognitive Interventions for M-FEI

Current metacognitive interventions do not take account of learners' metacognitive states completely, particularly instances of uncertainty (learners do not know their confidence) regarding performance. Given the significance of addressing such uncertainty, future research should prioritize the development of targeted interventions that respond to this specific learner condition, as identified through the M-FEI method.

Additionally, findings from the second user evaluation study indicated that pupils who had the tailored intervention using M-FEI achieved significantly higher mathematical learning outcomes. This underscores the benefits of having intervention and potential in fostering metacognitive skills for sustained learning outcomes. Future research should investigate optimal intervention schedules and develop strategies for providing long-term metacognitive support. This line of inquiry could benefit from longitudinal studies that evaluate the effectiveness of such interventions over extended periods, offering deeper insights into their potential to foster continuous learning and

adaptive skill development.

6.4.4 New ITS for Dynamics of Learner's Metacognition

In typical classroom environments, metacognitive monitoring often occurs spontaneously, without explicit prompting from educational tools. This raises a critical question: Can expert teachers recognize instances of unstructured self-reflection through observable changes in a student's demeanor or facial expressions, particularly when students persist despite limited understanding?

For example, when students advance to new content without fully grasping preceding material, this behavior may reflect a form of imprecise metacognitive monitoring. In such cases, expert teachers are often able to detect these gaps and provide timely interventions to help students become aware of their misunderstandings, even in the absence of formal reflective cues.

Future research should focus on capturing data that replicates this intuitive expertise demonstrated by teachers. Such data would enable ITSs to emulate expert observations and deliver dynamic, context-sensitive support for learners' metacognitive monitoring. Analyzing spontaneous instances of metacognitive activity in naturalistic learning environments may yield critical insights into learner self-reflection processes and inform the design of more adaptive and responsive educational technologies.

6.5 Looking Ahead

Metacognitive skills develop rapidly in children between the ages of 7 and 11, a critical period when they begin to reflect on their own thinking, assess their knowledge, set learning goals, and make informed decisions about what to do next. Supporting the development of these skills is essential across all educational contexts, whether in traditional classrooms or CBLEs.

Fostering children's metacognitive skills should be recognized as a core responsibility of both educators and educational system designers. This PhD research has illustrated that AI-based techniques can effectively support this goal by interpreting non-verbal behaviors to adaptively provide metacognitive guidance, much like expert teachers do when evaluating their students' self-reflections. Importantly, these AI-driven interventions can be integrated into learning environments without com-

promising learners' privacy, offering timely and individualized support that enhances both learning outcomes and the development of critical thinking skills. Moreover, this innovation offers particular benefits for learners who face additional challenges in learning or participation, as the AI-driven system provides unlimited patience and consistent support, which helps to deliver significant and sustained improvements in their educational experiences.

The contributions of this research not only alleviate some of the pressures faced by teachers in cultivating metacognitive skills but also offer them a complementary tool to better support children's learning journeys.

Looking forward, as CBLEs become increasingly prevalent and accessible, the integration of AI-driven metacognitive support systems holds significant promise for transforming educational practices on a broader scale. Future work will expand this research by exploring multi-modal data sources, including audio signals and contextual behavioral cues, to provide even more nuanced and effective support for learners. Additionally, these technologies will be extended to support a wider range of learning contexts beyond mathematics, including language acquisition, science education, and other subject areas where metacognitive skills play a vital role. Ultimately, this line of inquiry aspires to make personalized, metacognitively aware learning experiences available to all children, empowering them to become more independent, reflective, and capable learners in an increasingly complex world.

References

Abdelshiheed, M., Hostetter, J. W., Barnes, T. & Chi, M. (2023), Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems, *in* 'International Conference on Artificial Intelligence in Education', Springer, pp. 291–303.

Abdelshiheed, M., Hostetter, J. W., Shabrina, P., Barnes, T. & Chi, M. (2023), 'The power of nudging: Exploring three interventions for metacognitive skills instruction across intelligent tutoring systems', *arXiv preprint arXiv:2303.11965* .

Ainley, M. (2006), 'Connecting with learning: Motivation, affect and cognition in interest processes', *Educational psychology review* **18**, 391–405.

Aly, M. (2024), 'Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model', *Multimedia Tools and Applications* pp. 1–40.

Amershi, S., Cakmak, M., Knox, W. B. & Kulesza, T. (2014), 'Power to the people: The role of humans in interactive machine learning', *AI magazine* **35**(4), 105–120.

Argyle, M., Cook, M. & Cramer, D. (1994), 'Gaze and mutual gaze', *The British Journal of Psychiatry* **165**(6), 848–850.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. & Schmid, C. (2021), Vivit: A video vision transformer, *in* 'Proceedings of the IEEE/CVF international conference on computer vision', pp. 6836–6846.

Azevedo, R., Bouchet, F., Duffy, M., Harley, J., Taub, M., Trevors, G., Cloude, E., Dever, D., Wiedbusch, M., Wortha, F. et al. (2022), 'Lessons learned and future

directions of metatutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system', *Frontiers in Psychology* **13**, 813632.

Azevedo, R. & Hadwin, A. F. (2005), 'Scaffolding self-regulated learning and metacognition–implications for the design of computer-based scaffolds', *Instructional science* **33**(5/6), 367–379.

Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C. & Fike, A. (2009), Metatutor: A metacognitive tool for enhancing self-regulated learning, *in* '2009 AAAI Fall symposium series'.

Baker, R. & Ocumpaugh, J. (2014), '16. interaction-based affect detection in educational software', *The Oxford handbook of affective computing* p. 233.

Baltaci, S. & Gokcay, D. (2016), 'Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features', *International Journal of Human–Computer Interaction* **32**(12), 956–966.

Baltrusaitis, T., Zadeh, A., Lim, Y. C. & Morency, L.-P. (2018), Openface 2.0: Facial behavior analysis toolkit, *in* '2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)', IEEE, pp. 59–66.

Barzilai, S. & Blau, I. (2014), 'Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences', *Computers & Education* **70**, 65–79.

Behera, A., Matthew, P., Keidel, A., Vangorp, P., Fang, H. & Canning, S. (2020), 'Associating facial expressions and upper-body gestures with learning tasks for enhancing intelligent tutoring systems', *International Journal of Artificial Intelligence in Education* **30**, 236–270.

Bellon, E., Fias, W. & De Smedt, B. (2020), 'Metacognition across domains: Is the association between arithmetic and metacognitive monitoring domain-specific?', *PLoS One* **15**(3), e0229932.

Benjamin, A. S. & Tullis, J. (2010), 'What makes distributed practice effective?', *Cognitive psychology* **61**(3), 228–247.

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L. & Rellinger, E. R. (1995), 'Metacognition and problem solving: A process-oriented approach.', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21**(1), 205.

Bertasius, G., Wang, H. & Torresani, L. (2021), Is space-time attention all you need for video understanding?, *in* 'ICML', Vol. 2, p. 4.

Bouma, G. (2009), 'Normalized (pointwise) mutual information in collocation extraction', *Proceedings of GSCL* **30**, 31–40.

Bradley, M. M., Greenwald, M. K., Petry, M. C. & Lang, P. J. (1992), 'Remembering pictures: pleasure and arousal in memory.', *Journal of experimental psychology: Learning, Memory, and Cognition* **18**(2), 379.

Brookman-Byrne, A., Mareschal, D., Tolmie, A. K. & Dumontheil, I. (2018), 'Inhibitory control and counterintuitive science and maths reasoning in adolescence', *PLoS One* **13**(6), e0198973.

Brosnan, M., Johnson, H., Grawemeyer, B., Chapman, E., Antoniadou, K. & Hollinworth, M. (2016), 'Deficits in metacognitive monitoring in mathematics assessments in learners with autism spectrum disorder', *Autism* **20**(4), 463–472.

Brown, G. T., Andrade, H. L. & Chen, F. (2015), 'Accuracy in student self-assessment: directions and cautions for research', *Assessment in Education: Principles, Policy & Practice* **22**(4), 444–457.

Bullen, J. C., Lerro, L. S., Zajic, M., McIntyre, N. & Mundy, P. (2020), 'A developmental study of mathematics in children with autism spectrum disorder, symptoms of attention deficit hyperactivity disorder, or typical development', *Journal of autism and developmental disorders* **50**(12), 4463–4476.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M. & Zisserman, A. (2018), Vggface2: A dataset for recognising faces across pose and age, *in* '2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)', IEEE, pp. 67–74.

Carpenter, K. L., Williams, D. M. & Nicholson, T. (2019), 'Putting your money where your mouth is: examining metacognition in asd using post-decision wagering', *Journal of autism and developmental disorders* **49**(10), 4268–4279.

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C. & Zisserman, A. (2018), 'A short note about kinetics-600', *arXiv preprint arXiv:1808.01340* .

Chen, L., Chen, P. & Lin, Z. (2020), 'Artificial intelligence in education: A review', *Ieee Access* **8**, 75264–75278.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P. & Glaser, R. (1989), 'Self-explanations: How students study and use examples in learning to solve problems', *Cognitive science* **13**(2), 145–182.

Chi, M. T., Siler, S. A. & Jeong, H. (2004), 'Can tutors monitor students' understanding accurately?', *Cognition and instruction* **22**(3), 363–387.

Clabaugh, C., Mahajan, K., Jain, S., Pakkar, R., Becerra, D., Shi, Z., Deng, E., Lee, R., Ragusa, G. & Matarić, M. (2019), 'Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders', *Frontiers in Robotics and AI* **6**, 110.

Cloude, E. B., Wortha, F., Dever, D. A. & Azevedo, R. (2020), How do emotions change during learning with an intelligent tutoring system? metacognitive monitoring and performance with metatutor., *in* 'CogSci'.

Cogliano, M., Bernacki, M. L. & Kardash, C. M. (2020), 'A metacognitive retrieval practice intervention to improve undergraduates' monitoring and control processes and use of performance feedback for classroom learning.', *Journal of Educational Psychology* .

Cohen, J. (2013), *Statistical power analysis for the behavioral sciences*, routledge.

Cosentino, G., Lee-Cultura, S., Papavlasopoulou, S. & Giannakos, M. (2023), Designing multi sensory environments for children' s learning: An analysis of teachers' and researchers' perspectives, *in* 'Proceedings of the 22nd Annual ACM Interaction Design and Children Conference', pp. 388–396.

Crawford, J. R., Stewart, L. & Moore, J. (1989), 'Demonstration of savings on the avlt and development of a parallel form', *Journal of Clinical and Experimental Neuropsychology* **11**(6), 975–981.

Desoete, A. & De Craene, B. (2019), 'Metacognition and mathematics education: An overview', *ZDM* **51**(4), 565–575.

Dignath, C. & Büttner, G. (2008), 'Components of fostering self-regulated learning among students. a meta-analysis on intervention studies at primary and secondary school level', *Metacognition and learning* **3**, 231–264.

Dignath, C. & Büttner, G. (2018), 'Teachers' direct and indirect promotion of self-regulated learning in primary and secondary school mathematics classes–insights from video-based classroom observations and teacher interviews', *Metacognition and Learning* **13**, 127–157.

Dimberg, U. & Lundquist, L.-O. (1990), 'Gender differences in facial reactions to facial expressions', *Biological psychology* **30**(2), 151–159.

Dismuke, C. & Lindrooth, R. (2006), 'Ordinary least squares', *Methods and designs for outcomes research* **93**(1), 93–104.

D'Mello, S. & Graesser, A. (2013), 'Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back', *ACM Transactions on Interactive Intelligent Systems (TiiS)* **2**(4), 1–39.

D'Mello, S., Graesser, A. et al. (2007), Monitoring affective trajectories during complex learning, *in* 'Proceedings of the annual meeting of the cognitive science society', Vol. 29.

D'Mello, S. K., Craig, S. D. & Graesser, A. C. (2009), 'Multimethod assessment of affective experience and expression during deep learning', *International Journal of Learning Technology* **4**(3-4), 165–187.

Dominguez-Catena, I., Paternain, D. & Galar, M. (2022), 'Assessing demographic bias transfer from dataset to model: A case study in facial expression recognition', *arXiv preprint arXiv:2205.10049* .

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. & Darrell, T. (2015), Long-term recurrent convolutional networks for visual recognition and description, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2625–2634.

Dubey, S. R., Singh, S. K. & Chaudhuri, B. B. (2022), 'Activation functions in deep learning: A comprehensive survey and benchmark', *Neurocomputing* **503**, 92–108.

Duff, K., Beglinger, L. J., Moser, D. J., Schultz, S. K. & Paulsen, J. S. (2010), 'Practice effects and outcome of cognitive training: Preliminary evidence from a memory training course', *The American Journal of Geriatric Psychiatry* **18**(1), 91.

D' Mello, S. & Graesser, A. (2012), 'Dynamics of affective states during complex learning', *Learning and Instruction* **22**(2), 145–157.

Ebbinghaus, H. (2013), '[image] memory: A contribution to experimental psychology', *Annals of neurosciences* **20**(4), 155.

Eberhart, J., Schäfer, F. & Bryce, D. (2025), 'Are metacognition interventions in young children effective? evidence from a series of meta-analyses', *Metacognition and Learning* **20**(1), 7.

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M. & Bunge, S. A. (2017), 'Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?', *Developmental cognitive neuroscience* **25**, 69–91.

Efklides, A. (2011), 'Interactions of metacognition with motivation and affect in self-regulated learning: The masrl model', *Educational psychologist* **46**(1), 6–25.

Ekman, P. (1999), 'Basic emotions', *Handbook of cognition and emotion* **98**(45-60), 16.

Ekman, P. & Friesen, W. V. (1978), 'Facial action coding system', *Environmental Psychology & Nonverbal Behavior* .

Ekman, P., Friesen, W. V. & Hager, J. C. (2002), *Facial Action Coding System: Manual and Investigator's Guide*, Salt Lake City, UT. Available from the Paul Ekman Group: `https://www.paulekman.com/facial-action-coding-system/`.

EPCC (2025), 'Cirrus', `https://www.epcc.ed.ac.uk/hpc-services/cirrus`. [Online; accessed 10-September-2025].

Epley, N. & Gilovich, T. (2006), 'The anchoring-and-adjustment heuristic: Why the adjustments are insufficient', *Psychological science* **17**(4), 311–318.

Espinosa, L., Arciniegas, A., Cortes, Y., Prieto, F. & Brancheriau, L. (2017), 'Automatic segmentation of acoustic tomography images for the measurement of wood decay', *Wood Science and Technology* **51**, 69–84.

FaceReader (5.0) (2025). Available from Noldus Information Technology.

Fanari, R., Meloni, C. & Massidda, D. (2019), 'Visual and spatial working memory abilities predict early math skills: A longitudinal study', *Frontiers in Psychology* **10**, 2460.

Fang, H. L. (2012), *The effects of simplified schema-based instruction on elementary students' mathematical word problem solving performance*, The University of Southern Mississippi.

Fauvel, S., Yu, H., Miao, C., Cui, L., Song, H., Zhang, L., Li, X. & Leung, C. (2018), Artificial intelligence powered moocs: a brief survey, *in* '2018 IEEE international conference on agents (ICA)', IEEE, pp. 56–61.

Feichtenhofer, C., Fan, H., Malik, J. & He, K. (2019), Slowfast networks for video recognition, *in* 'Proceedings of the IEEE/CVF international conference on computer vision', pp. 6202–6211.

Field, A. (2013), *Discovering Statistics Using IBM SPSS Statistics*, Sage.

Flavell, J. H. (1979), 'Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry.', *American psychologist* **34**(10), 906.

Flores, N. & Lewis, M. (2023), ' "false positives, re-entry programs and long term english learners" : Undoing dichotomous frames in us language education policy', *Equity & Excellence in Education* pp. 1–13.

German Research Center for Artificial Intelligence (DFKI) (2025), `https://www.dfki.de/en/web`. Accessed: 2025-03-28.

github/affect2mmp (2025), 'Affect to metacognition networks', `https://github.com/XRR422/Affect2Metacognition`.

Graesser, A. C. (2020), 'Emotions are the experiential glue of learning environments in the 21st century', *Learning and Instruction* **70**, 101212.

Graesser, A., Chipman, P., King, B., McDaniel, B. & D'Mello, S. (2007), 'Emotions and learning with auto tutor', *Frontiers in Artificial Intelligence and Applications* **158**, 569.

Grainger, C., Williams, D. M. & Lind, S. E. (2016), 'Metacognitive monitoring and control processes in children with autism spectrum disorder: Diminished judgement of confidence accuracy', *Consciousness and Cognition* **42**, 65–74.

Grawemeyer, B., Johnson, H. & Brosnan, M. (2015), Can young people with autism spectrum disorder benefit from an open learner model?, *in* 'International Conference on Artificial Intelligence in Education', Springer, pp. 591–594.

Great Britain (2018), 'Data protection act 2018', `https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted`. Accessed: 2024-11-19.

Guo, W. (2020), Using metacognitive monitoring feedback to improve student learning in augmented reality environments, PhD thesis, University of Missouri-Columbia.

Gupta, A., D'Cunha, A., Awasthi, K. & Balasubramanian, V. (2016), 'Daisee: Towards user engagement recognition in the wild', *arXiv preprint arXiv:1609.01885* .

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J. & Seung, H. S. (2000), 'Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit', *nature* **405**(6789), 947–951.

Halmo, S. M., Yamini, K. A. & Stanton, J. D. (2024), 'Metacognition and self-efficacy in action: How first-year students monitor and use self-coaching to move past metacognitive discomfort during problem solving', *CBE—Life Sciences Education* **23**(2), ar13.

Hamrouni, A. & Bendella, F. (2024), 'Recognizing students emotions in game-based learning environment', *International Journal of Information Technology* pp. 1–11.

Harley, J. M., Bouchet, F., Hussain, M. S., Azevedo, R. & Calvo, R. (2015), 'A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system', *Computers in Human Behavior* **48**, 615–625.

Harris, L. R. & Brown, G. T. (2013), 'Opportunities and obstacles to consider when using peer-and self-assessment to improve student learning: Case studies into teachers' implementation', *Teaching and Teacher Education* **36**, 101–111.

Harter, S. (2012), 'Developmental differences in self-representations during childhood', *The construction of the self: Developmental and sociocultural foundations* pp. 27–71.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. (1998), 'Support vector machines', *IEEE Intelligent Systems and their applications* **13**(4), 18–28.

Herzog, M. & Casale, G. (2022), 'The effects of a computer-based mathematics intervention in primary school students with and without emotional and behavioral difficulties', *International Electronic Journal of Elementary Education* **14**(3), 303–317.

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A., Coleman, R., Henderson, P., Major, L., Coe, R. & Mason, D. (2016), 'The sutton trust-education endowment foundation teaching and learning toolkit.'.

Hugging Face (2025), 'Hugging Face –The AI community building the future'. Accessed: 2025-05-15.
URL: *https://huggingface.co/*

iMotions (2018), 'Attention tool', `https://imotions.com/`. [Online; accessed 28-December-2022].

Isaacson, R. M. & Fujita, F. (2006), 'Metacognitive knowledge monitoring and self-regulated learning: Academic success and reflections on learning.', *Journal of Scholarship of Teaching and Learning* **6**(1), 39–55.

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R. & Schyns, P. G. (2012), 'Facial expressions of emotion are not culturally universal', *Proceedings of the National Academy of Sciences* **109**(19), 7241–7244.

Jitendra, A. K. & Star, J. R. (2011), 'Meeting the needs of students with learning disabilities in inclusive mathematics classrooms: The role of schema-based instruction on mathematical problem-solving', *Theory into practice* **50**(1), 12–19.

Jo, E. S. & Gebru, T. (2020), Lessons from archives: Strategies for collecting sociocultural data in machine learning, *in* 'Proceedings of the 2020 conference on fairness, accountability, and transparency', pp. 306–316.

Jowett, E., Moore, D. W. & Anderson, A. (2012), 'Using an ipad-based video modelling package to teach numeracy skills to a child with an autism spectrum disorder', *Developmental neurorehabilitation* **15**(4), 304–312.

Kautzmann, T. R., Carlotto, T. & Jaques, P. A. (2016), Adaptive training of the metacognitive skill of knowledge monitoring in intelligent tutoring systems, *in* 'Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings 13', Springer, pp. 301–306.

Kautzmann, T. R. & Jaques, P. A. (2019), 'Effects of adaptive training on metacognitive knowledge monitoring ability in computer-based learning', *Computers & Education* **129**, 92–105.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P. et al. (2017), 'The kinetics human action video dataset', *arXiv preprint arXiv:1705.06950* .

Kelleher, C. & Hnin, W. (2019), Predicting cognitive load in future code puzzles, *in* 'Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems', pp. 1–12.

Kim, J. (2024), 'Leading teachers' perspective on teacher-ai collaboration in education', *Education and Information Technologies* **29**(7), 8693–8724.

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016), 'Inherent trade-offs in the fair determination of risk scores', *arXiv preprint arXiv:1609.05807* .

Komatani, K. & Nakano, M. (2020), User impressions of questions to acquire lexical knowledge, *in* 'Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue', pp. 147–156.

Koriat, Lichtenstein, . F. (1981), 'Journal of experimental psychology. human learning and memory', *American Psychological Association* .

Kort, B., Reilly, R. & Picard, R. W. (2001), An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion, *in* 'Proceedings IEEE international conference on advanced learning technologies', IEEE, pp. 43–46.

Leavy, S., Siapera, E. & O'Sullivan, B. (2021), Ethical data curation for ai: An approach based on feminist epistemology and critical theories of race, *in* 'Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society', pp. 695–703.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *nature* **521**(7553), 436–444.

Lehnert, F. K. (2024), 'Measuring children' s user experience with e-assessments: Implications for a better interpretation of ux evaluation methods for school-aged children'.

Lepper, M. R. & Henderlong, J. (2000), 'Turning "play" into "work" and "work" into "play" : 25 years of research on intrinsic versus extrinsic motivation', *Intrinsic and extrinsic motivation* pp. 257–307.

Li, H., Hua, X., Yang, Y., Huang, B. & Si, J. (2020), 'How does task switching affect arithmetic strategy use in children with low mathematics achievement? evidence from computational estimation', *European Journal of Psychology of Education* **35**, 225–240.

Li, S. & Deng, W. (2020), 'Deep facial expression recognition: A survey', *IEEE transactions on affective computing* .

Lindsley, O. R. (1991), 'Precision teaching's unique legacy from bf skinner', *Journal of Behavioral Education* **1**, 253–266.

Linson, A., Xu, Y., English, A. R. & Fisher, R. B. (2022), Identifying student struggle by analyzing facial movement during asynchronous video lecture viewing: Towards an automated tool to support instructors, *in* 'International Conference on Artificial Intelligence in Education', Springer, pp. 53–65.

Liu, J., McKenna, T. M., Gribok, A., Beidleman, B. A., Tharion, W. J. & Reifman, J. (2008), 'A fuzzy logic algorithm to assign confidence levels to heart and respiratory rate time series', *Physiological measurement* **29**(1), 81.

Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J. & Mei, T. (2020), Learning to localize actions from moments, *in* 'Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16', Springer, pp. 137–154.

Mandler, G. (1989), Affect and learning: Causes and consequences of emotional interactions, *in* 'Affect and mathematical problem solving: A new perspective', Springer, pp. 3–19.

Maras, K., Gamble, T. & Brosnan, M. (2019), 'Supporting metacognitive monitoring in mathematics learning for young people with autism spectrum disorder: A classroom-based study', *Autism* **23**(1), 60–70.

Massoli, F. V., Amato, G. & Falchi, F. (2020), 'Cross-resolution learning for face recognition', *Image and Vision Computing* **99**, 103927.

McAlenney, A. L. & Coyne, M. D. (2015), 'Addressing false positives in early reading assessment using intervention response data', *Learning Disability Quarterly* **38**(1), 53–65.

Mevarech, Z. R. & Kramarski, B. (1997), 'Improve: A multidimensional method for teaching mathematics in heterogeneous classrooms', *American educational research journal* **34**(2), 365–394.

Mihalca, L. & Mengelkamp, C. (2020), 'Effects of induced levels of prior knowledge on monitoring accuracy and performance when learning from self-regulated problem solving.', *Journal of Educational Psychology* **112**(4), 795.

Montero, N. A. et al. (2021), 'The impact of a metacognitive intervention using improve model on grade 7 students' metacognitive awareness in mathematics', *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* **12**(3), 3881–3894.

Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S. & Ling, H. (2022), Expanding language-image pretrained models for general video recognition, *in* 'European Conference on Computer Vision', Springer, pp. 1–18.

Ohtani, K. & Hisasaka, T. (2018), 'Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance', *Metacognition and Learning* **13**, 179–212.

ONNX Community (2025), 'Open neural network exchange (onnx)', `https://onnx.ai/`. Accessed: 2025-04-21.

OpenAI (2025), *OpenAI API Documentation*. Accessed: 2025-03-07.
URL: *https://platform.openai.com/docs/api-reference/introduction*

Orji, F. A. & Vassileva, J. (2022), 'Automatic modeling of student characteristics with interaction and physiological data using machine learning: A review', *Frontiers in Artificial Intelligence* **5**, 1015660.

Ortegano, L. & Ramírez, E. (2019), 'Serious educational reinforcement game in preschool', *arXiv preprint arXiv:1909.10337* .

Oviatt, S. (2006), Human-centered design meets cognitive load theory: designing interfaces that help people think, *in* 'Proceedings of the 14th ACM international conference on Multimedia', pp. 871–880.

Oviatt, S., Cohen, P., Wu, L., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J. et al. (2000), 'Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions', *Human-computer interaction* **15**(4), 263–322.

Pekrun, R. & Linnenbrink-Garcia, L. (2012), Academic emotions and student engagement, *in* 'Handbook of research on student engagement', Springer, pp. 259–282.

Peng, Y., Zhao, Y. & Zhang, J. (2018), 'Two-stream collaborative learning with spatial-temporal attention for video classification', *IEEE Transactions on Circuits and Systems for Video Technology* **29**(3), 773–786.

Person, N. K., Graesser, A. C., Magliano, J. P. & Kreuz, R. J. (1994), 'Inferring what the student knows in one-to-one tutoring: The role of student questions and answers', *Learning and individual differences* **6**(2), 205–229.

Pescetelli, N. & Yeung, N. (2021), 'The role of decision confidence in advice-taking and trust formation.', *Journal of Experimental Psychology: General* **150**(3), 507.

Piaget, J. (1952), 'The origins of intelligence in children', *International University* .

Pintrich, P. R., Wolters, C. A. & Baxter, G. P. (2000), '2. assessing metacognition and self-regulated learning'.

Pourmirzaei, M., Montazer, G. A. & Mousavi, E. (2023), 'Attendee: an affective tutoring system based on facial emotion recognition and head pose estimation to personalize e-learning environment', *Journal of Computers in Education* pp. 1–28.

Psaltakis, G., Rogdakis, K., Loizos, M. & Kymakis, E. (2024), 'One-vs-one, one-vs-rest, and a novel outcome-driven one-vs-one binary classifiers enabled by optoelectronic memristors towards overcoming hardware limitations in multiclass classification', *Discover Materials* **4**(1), 7.

Qiao, H., Tan, J., Wen, S., Zhang, M., Xu, S. & Jin, L. (2024), 'De novo dissecting the three-dimensional facial morphology of 2379 han chinese individuals', *Phenomics* **4**(1), 1–12.

Riku, A. (2021), 'Mindless attractor: A false-positive resistant intervention for drawing attention using auditory perturbation. in chi'21. acm, new york', *NY* **99**, 1.

Rivers, M. L., Dunlosky, J. & Persky, A. M. (2020), 'Measuring metacognitive knowledge, monitoring, and control in the pharmacy classroom and experiential settings', *American journal of pharmaceutical education* **84**(5), 7730.

Romesburg, C. (2004), *Cluster analysis for researchers*, Lulu. com.

Ruan, X., Constantin, A., Palansuriya, C., Wang, K. & Atkinson, M. (2025), Nurturing self-aware learning through facial expression interpretation, *in* 'Proceedings of the CHI Conference on Human Factors in Computing Systems'.

Ruan, X., Palansuriya, C. & Constantin, A. (2022), Real-time feedback based on emotion recognition for improving children' s metacognitive monitoring skill, *in* 'Interaction Design and Children', pp. 672–675.

Ruan, X., Palansuriya, C., Constantin, A. & Tsiakas, K. (2023), Supporting children' s metacognition with a facial emotion recognition based intelligent tutor system, *in* 'Proceedings of the 22nd Annual ACM Interaction Design and Children Conference', pp. 502–506.

Rudovic, O., Lee, J., Dai, M., Schuller, B. & Picard, R. W. (2018), 'Personalized machine learning for robot perception of affect and engagement in autism therapy', *Science Robotics* **3**(19), eaao6760.

Rudovic, O., Utsumi, Y., Lee, J., Hernandez, J., Ferrer, E. C., Schuller, B. & Picard, R. W. (2018), Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism, *in* '2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)', IEEE, pp. 339–346.

Samizadeh, S. (2022), 'Characteristics of asian faces', *Non-Surgical Rejuvenation of Asian Faces* pp. 41–58.

Sawyer, A. C., Williamson, P. & Young, R. (2014), 'Metacognitive processes in emotion recognition: Are they different in adults with asperger' s disorder?', *Journal of Autism and Developmental Disorders* **44**(6), 1373–1382.

Schraw, G. (2009), 'A conceptual analysis of five measures of metacognitive monitoring', *Metacognition and learning* **4**, 33–45.

Schraw, G. & Dennison, R. S. (1994), 'Assessing metacognitive awareness', *Contemporary educational psychology* **19**(4), 460–475.

Shandong Education Press (2025), `https://www.sjs.com.cn/en/Index.html`. Accessed: 2025-03-24.

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C. & Frith, C. D. (2014), 'Supra-personal cognitive control and metacognition', *Trends in cognitive sciences* **18**(4), 186–193.

Shi, Z., Groechel, T. R., Jain, S., Chima, K., Rudovic, O. & Matarić, M. J. (2021), 'Toward personalized affect-aware socially assistive robot tutors in long-term interventions for children with autism', *arXiv preprint arXiv:2101.10580* .

Simmons, J. P., LeBoeuf, R. A. & Nelson, L. D. (2010), 'The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors?', *Journal of personality and social psychology* **99**(6), 917.

Singh, M., Hoque, X., Zeng, D., Wang, Y., Ikeda, K. & Dhall, A. (2023), Do i have your attention: A large scale engagement prediction dataset and baselines,

in 'Proceedings of the 25th International Conference on Multimodal Interaction', pp. 174–182.

Siregar, N. C., Rosli, R., Maat, S. M., Alias, A., Toran, H., Mottan, K. & Nor, S. M. (2020), 'The impacts of mathematics instructional strategy on students with autism: A systematic literature review.', *European Journal of Educational Research* **9**(2), 729–741.

Song, M.-K. & Ward, S. E. (2015), 'Assessment effects in educational and psychosocial intervention trials: An important but often-overlooked problem', *Research in nursing & health* **38**(3), 241–247.

Sperling, R. A., Richmond, A. S., Ramsay, C. M. & Klapp, M. (2012), 'The measurement and predictive ability of metacognition in middle school learners', *The Journal of Educational Research* **105**(1), 1–7.

Stehman, S. V. (1997), 'Selecting and interpreting measures of thematic classification accuracy', *Remote sensing of Environment* **62**(1), 77–89.

Sundararajan, M., Taly, A. & Yan, Q. (2017), Axiomatic attribution for deep networks, *in* 'International conference on machine learning', PMLR, pp. 3319–3328.

Taub, M. & Azevedo, R. (2018), 'Using sequence mining to analyze metacognitive monitoring and scientific inquiry based on levels of efficiency and emotions during game-based learning.', *Journal of Educational Data Mining* **10**(3), 1–26.

Taub, M., Azevedo, R. & Mudrick, N. V. (2018), How do different levels of au4 impact metacognitive monitoring during learning with intelligent tutoring systems?, *in* 'Intelligent Tutoring Systems: 14th International Conference, ITS 2018, Montreal, QC, Canada, June 11–15, 2018, Proceedings 14', Springer, pp. 223–232.

Taub, M., Azevedo, R., Rajendran, R., Cloude, E. B., Biswas, G. & Price, M. J. (2021), 'How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system?', *Learning and Instruction* **72**, 101200.

Tisza, G., Sharma, K., Papavlasopoulou, S., Markopoulos, P. & Giannakos, M. (2022), Understanding fun in learning to code: A multi-modal data approach, *in* 'Interaction Design and Children', pp. 274–287.

Tombazzi, A., Choukeir, J. & Lai, N. (2023), 'The shifting landscape of learning and ai', *RSA Blog* .
URL: *https://www.thersa.org/blog/2023/08/ai-learning-shifting-landscape*

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023), 'Llama: Open and efficient foundation language models', *arXiv preprint arXiv:2302.13971* .

Träff, U., Olsson, L., Skagerlund, K., Skagenholt, M. & Östergren, R. (2019), 'Logical reasoning, spatial processing, and verbal working memory: Longitudinal predictors of physics achievement at age 12–13 years', *Frontiers in psychology* **10**, 1929.

Tsamago, H. E. & Bayaga, A. (2024), 'Investigating the correlation between metacognitive skills and conceptual understanding using self-organised learning environments pedagogy', *South African Journal of Education* **44**(3).

Tsiakas, K., Barakova, E., Khan, J. V. & Markopoulos, P. (2020), Brainhood: towards an explainable recommendation system for self-regulated cognitive training in children, *in* 'Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments', pp. 1–6.

Urban, K. & Urban, M. (2021*a*), 'Anchoring effect of performance feedback on accuracy of metacognitive monitoring in preschool children', *Europe's journal of psychology* **17**(1), 104.

Urban, K. & Urban, M. (2021*b*), 'Effects of performance feedback and repeated experience on self-evaluation accuracy in high-and low-performing preschool children', *European Journal of Psychology of Education* **36**(1), 109–124.

Valencia, K., Rusu, C., Quiñones, D. & Jamet, E. (2019), 'The impact of technology on people with autism spectrum disorder: a systematic literature review', *Sensors* **19**(20), 4485.

Vanneste, P., Oramas, J., Verelst, T., Tuytelaars, T., Raes, A., Depaepe, F. & Van den Noortgate, W. (2021), 'Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement', *Mathematics* **9**(3), 287.

Wang, T., Zheng, J., Tan, C. & Lajoie, S. P. (2023), 'Computer-based scaffoldings influence students' metacognitive monitoring and problem-solving efficiency in an intelligent tutoring system', *Journal of Computer Assisted Learning* **39**(5), 1652–1665.

Wang, Y. & Sperling, R. A. (2020), Characteristics of effective self-regulated learning interventions in mathematics classrooms: A systematic review, *in* 'Frontiers in Education', Vol. 5, Frontiers Media SA, p. 58.

Winne, P. H. (2011), 'A cognitive and metacognitive analysis of self-regulated learning', *Handbook of self-regulation of learning and performance* pp. 15–32.

Winne, P. H. & Hadwin, A. F. (2010), 'Self-regulated learning and socio-cognitive theory', *International encyclopedia of education* pp. 503–508.

Wojcik, D. Z., Moulin, C. J. & Souchay, C. (2013), 'Metamemory in children with autism: Exploring "feeling-of-knowing" in episodic and semantic memory.', *Neuropsychology* **27**(1), 19.

Wood, H. & Wood, D. (1999), 'Help seeking, learning and contingent tutoring', *Computers & Education* **33**(2-3), 153–169.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F. et al. (2024), 'Qwen2 technical report', *arXiv preprint arXiv:2407.10671* .

Yonglan, L., Huixin, Y., Xinghua, Z., Keli, Y., Jinping, B. & Lianbin, Z. (2023), 'Head and facial features of populations in different geographical regions of china', *Acta Anthropologica Sinica* **42**(06), 793.

Zumbach, J., Rammerstorfer, L. & Deibl, I. (2020), 'Cognitive and metacognitive support in learning with a serious game about demographic change', *Computers in Human Behavior* **103**, 120–129.