

HEC MONTRÉAL

**Neurophysiological Effects of AI-Mediated Mathematics Education: Exploring EEG
Measures of Task Load and the User's Experience**

par

Uriel Karerwa

Alexander J. Karran

HEC Montréal

Codirecteur de recherche

Sylvain Sénécal

HEC Montréal

Codirecteur de recherche

Sciences de la gestion

(User Experience)

*Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion (M. Sc.)*

Dec 2025

© Uriel Karerwa, 2025

Ethics Approval Form



Comité d'éthique de la recherche

Le 23 avril 2025

À l'attention de : Pierre-Majorique Léger, Technologies de l'information, HEC Montréal

Cochercheurs : Sylvain Sénécal; Patrick Charland; Bella Tadson; Noémie Beauchemin; François Courtemanche; David Briegne; Amine Abdessemed; Salima Tazi; Alexander John Karran; Jared Boasen; Ioana Mihaela Stefanescu; Katrina Sollazzo; Thaddé Rolon-Mérette; Alejandra Ruiz Segura; Uriel Kärerwa; Hajar Laghmari

Projet # : 2023-5071

Titre du projet : Interfaces neuro-adaptatives en fonction de la charge cognitive dans un contexte d'apprentissage en ligne et de magasinage en ligne.

Bonjour Pierre-Majorique Léger,

Pour donner suite à l'évaluation de votre formulaire F8 - Modification de projet, le CER de HEC Montréal vous informe de sa décision :

Les modifications ont été approuvées et notées au dossier. Le certificat actuel demeure valide jusqu'au prochain renouvellement.

Date d'approbation du projet : 13 juin 2022

Date d'échéance du certificat : 01 novembre 2025

En vous remerciant cordialement,

Le CER de HEC Montréal

Résumé

L'adoption rapide des systèmes de tutorat fondés sur l'intelligence artificielle a renforcé l'intérêt pour leur influence sur le processus d'apprentissage. Ce mémoire examine l'effet d'un tuteur en mathématiques basé sur l'IA, adapté au moyen d'une persona, sur l'effort cognitif, l'engagement et la compréhension à court terme des apprenants. La personnalisation s'appuie sur des profils issus du modèle des styles d'apprentissage de Felder et Silverman. Quarante participants ont complété trois séances de tutorat en mathématiques, alternant entre une version personnalisée et une version standard du tuteur. Des mesures neurophysiologiques, comportementales et issues de questionnaires ont été recueillies afin d'offrir un portrait global de l'expérience d'apprentissage.

La personnalisation n'a pas amélioré la précision aux questionnaires ni la satisfaction déclarée. Toutefois, les données physiologiques montrent que les apprenants déployaient moins d'effort mental lorsqu'ils interagissaient avec le tuteur personnalisé. Plus précisément, des diminutions du ratio θ frontal sur α pariétal ainsi qu'une réduction de la dilatation pupillaire indiquent un traitement plus efficace de l'information lorsque la communication est alignée sur les préférences des utilisateurs. Les comportements différaient également, les participants tentant un plus grand nombre de questions et utilisant plus fréquemment la notation mathématique lors des séances personnalisées. Ces résultats montrent qu'un ajustement du style communicationnel du tuteur peut influencer la manière dont les apprenants mobilisent leurs ressources cognitives, même en l'absence de gains immédiats de performance. Ils démontrent aussi la valeur de l'intégration d'indicateurs neurophysiologiques dans l'évaluation des systèmes de tutorat fondés

sur l'IA et soulignent comment les choix de conception liés aux persona peuvent façonner certains aspects de l'expérience des apprenants dans un contexte d'apprentissage médié par l'IA.

Mots clés : persona, personnalisation, apprentissage médié par l'IA, système de tutorat intelligent (ITS), charge cognitive, EEG, charge de tâche, apprentissage, engagement, satisfaction.

Méthodes de recherche : expérience quantitative en laboratoire utilisant des mesures physiologiques, dans un plan expérimental mixte ($2 \times 2 \times 3$).

Abstract

The rapid adoption of artificial intelligence tutoring systems has intensified interest in how these tools shape the learning process. This thesis evaluates the influence of a persona-adapted AI mathematics tutor on learners' cognitive effort, engagement, and understanding. Personalization was based on profiles from the Felder-Silverman Learning Style model. Forty participants completed three mathematics tutoring sessions, alternating between a personalized and a standard tutor. Neurophysiological, behavioural, and questionnaire-based measures were collected to provide a broad view of the learning experience. Personalization did not increase quiz accuracy or reported satisfaction. However, the physiological data indicated that learners expended less mental effort when interacting with the personalized tutor. Specifically, reductions in frontal theta to parietal alpha ratios and smaller pupil dilations, indicating reduced cognitive load and suggesting more efficient processing of instructional information, were observed with persona changes. Behavioural patterns also differed, with participants attempting more questions and incorporating mathematical notation more frequently in personalized sessions. The results demonstrate the value of integrating neurophysiological indicators into the evaluation of AI tutoring systems and highlight how persona design choices can shape aspects of the learners' experience in an AI-mediated learning context.

Keywords: Persona, Personalization, AI-mediated learning, Intelligent Tutoring System (ITS), Cognitive Load, EEG, Task Load, learning, engagement, satisfaction.

Research Methods: Quantitative lab experiment using physiological measurement tools, in a within-between (2X2X3) experimental design.

Table of contents

Ethics Approval Form.....	ii
Résumé.....	iii
Abstract.....	v
Table of contents.....	vii
List of tables and figures.....	xi
List of abbreviations and acronyms	xiii
Preface.....	xiv
Acknowledgements.....	xv
Chapter 1: Introduction.....	1
Thesis Structure	7
Personal Contribution	8
References.....	11
Chapter 2: Neurophysiological effects of AI-mediated mathematics education: Exploring EEG measures of task load and the user’s experience	14
Abstract.....	14
1. Introduction.....	16
2. Related Work	20
2.1 AI-Mediated Learning and Intelligent Tutoring Systems (ITS)	20
2.2 Theoretical Framework: Cognitive Load Theory	23

2.3 Theoretical Framework: Dual-Process Theory	24
2.4 Neurophysiological Indicators of Task Load.....	26
2.5 Felder-Silverman Learning Style Model	29
2.6 Learning Metrics and User Experience.....	32
3. Hypothesis Development	33
3.1 Learning Outcomes	34
3.2 Task Load.....	35
3.3 User Satisfaction	36
4. Methods.....	37
4.1 Participants.....	37
4.2 Study Design.....	38
4.3 Materials	39
4.4 Procedure	43
4.5 Data Acquisition	47
4.6 Preprocessing and Feature Extraction.....	48
4.7 Measures	49
4.8 Statistical Analysis.....	54
5. Results.....	55
5.1 Effects of Personalization	56
5.2 Learning Gain	57
5.3 Satisfaction.....	58
5.4 Task Load.....	58

5.5 Control Variables	61
5.6 Triangulation.....	62
5.7 AI Tutor Interactions.....	63
5.8 Research Model Overview.....	64
6. Discussion	66
6.1 Interpretation.....	66
6.2 Methodological Contributions: Multimodal Measurement	75
6.3 Practical Implications for AI-Mediated Education	76
6.4 Limitations and Future Directions	77
7. Conclusion	78
References.....	81
Chapter 3: Conclusion.....	87
Theoretical Contributions	88
Practical Implications.....	88
Future Research Directions.....	89
Concluding Remarks.....	90
References.....	91
Appendices.....	92
Appendix A: Participant Exclusion Criteria	92
Appendix B: Post-Task Math Assessments	93
Appendix C: System-Level Persona Prompts.....	94
Appendix D: Expanded Results Section	95

Appendix E: Generative AI Use Disclosure Table	97
--	----

List of tables and figures

Figure 1. Research Model and Conceptual Framework of Effects of Personalized AI-Mediated Learning on Task Load and User Experience.....	34
Figure 2. Laboratory Design Demonstration: Participant View, Lab Assistant Demonstration, and Observation Room Setup	44
Figure 3. Experimental Design Including Steps and Duration of Each Phase.....	44
Figure 4. Topographic Distribution of Theta Band Power During Tutoring Tasks.....	60
Figure 5. Average Task Load Across Conditions Measured by EEG and Pupillometry	61
Figure 6. Behavioural and Physiological Metrics Across Tasks in Chronological Order	64
Figure 7. Updated Research Model Incorporating Results from Hypothesis Testing	65
Table 1. Personal Contributions and Responsibilities Throughout the Research Project.....	8
Table 2. Participant Demographics: Age, Sex, and Education Level	37
Table 3. 2X2X3 Factorial Design with Between- and Within-Subjects Comparisons	39
Table 4. Learning Styles Applied in the Experiment and Prompt Segments Used to Personalize the AI Math Tutor	41
Table 5. Description of Constructs and Measurement Tools Along with Cronbach’s Alpha.....	52
Table 6. Hypothesis Test Results	55
Table 7. Linear Mixed-Effects Results Comparing FSLSM Personalization and Control Conditions	57
Table 8. Statistical Control Variable Screening Results from Linear Regression of Personalization Effects on EEG-Derived Task Load.....	61
Table 9. Correlation Results Between Satisfaction and Dependent Variables	62

Table 10 (Appendix A1). Inclusion and Exclusion Criteria for Test Participants	92
Table 11 (Appendix D1). Descriptive Statistics table for Felder Silverman Learner Subgroups (Group A vs. Group B)	95
Table 12 (Appendix D2). Repeated Measures ANOVA Results for Task Order Comparison... ..	95
Table 13 (Appendix D13). Results from Repeated Measures ANOVA Comparing Control and Groups A and B	96
Table 14 (Appendix E1). AGI Use Disclosure Table	97

List of abbreviations and acronyms

AI: Artificial Intelligence

BCI: Brain Computer Interface

CLT: Cognitive Load theory

CL: Cognitive Load

DPT: Dual Process Theory

FSLs: Felder-Silverman Learning Style

Gen AI: Generative Artificial Intelligence

ITS: Intelligent Tutoring System

LLM: Large Language Model

TL: Task Load

Preface

An authorization to write the following dissertation has been granted by the administrative direction of the Master of Science program specializing in user experience. This dissertation is written in the form of a journal article. The agreement of all co-authors for this article has been obtained and submitted along with my thesis.

Generative artificial intelligence tools were used during the preparation of this thesis for tasks such as grammar checking, sentence refinement, and the translation to French. All content produced with the assistance of these tools was reviewed, revised, and validated by the authors, who take full responsibility for the final manuscript. A complete description of the AI-assisted methods used in the writing process is provided in Appendix E.

In April 2025, the HEC Montréal Research Ethics Board (CER) approved the renewed research project (Certificate #2023-5071), as shown in page 2.

The article in chapter 2 examines the Neurophysiological effects of AI-mediated learning through EEG measured task load and by measuring the user's experience.

Acknowledgements

Science is deeply collaborative, and I could not have accomplished this work without the contributions of so many. To Hajar and Malayka, bonded through countless hours collecting data in the lab over the summer, it was an honor to work alongside you both. I would like to thank the Fundamental science team at the Tech3Lab for all the support and guidance throughout this journey namely Thadd , Katrina, Alejandra and my director of studies Alexander. I am also grateful to David, Elise, Salima, and Xavier for their tireless efforts on the operations side, helping get the experiment up and running and skillfully managing large fires as they arose. Thanks to Shang-Lin Chen for support with the statistical analysis and helping me verify my results. I also wouldn't be here without support from the great people I've had the privilege of working with in previous years like Dr. Reich and Dr. Kolb, thanks for your many recommendation letters.

To my wonderful partner Melissa, who stood by me during my move across the country and continued to support me through the many hardships of this period: I am profoundly grateful. To my extended family, thank you for helping me get settled during those crucial first few months of my move to Montreal and checking in to see how I was doing. Finally, to my Montreal flat mates and friends (Sonia, Kara and Caleigh) thanks for making it a safe and fun place to live. With all your help, I survived.

Chapter 1: Introduction

Education systems worldwide face increasing difficulty meeting fundamental learning needs.

UNESCO estimates that 617 million students worldwide fail to achieve minimum proficiency in learning (UNESCO, 2023). In Canada, one in five students now performs below the baseline proficiency level in mathematics, the highest proportion recorded since the country began participating in the Programme for International Student Assessment (OECD, 2023a). Similar declines are observed across many OECD countries, where mathematics scores have reached their lowest levels in two decades (OECD, 2023a, 2023b). These observed trends threaten long-term social mobility and preparedness for STEM-relevant occupations.

As school performance continues to plummet, learners increasingly turn to artificial intelligence systems to support their studies, yet little is known about how these tools can improve learning processes. According to a report released by the Pew Research Center (N = 1,391), the proportion of teens, In the United States (ages 13-17) who report using generative artificial intelligence (AI) for schoolwork doubled between 2023 and 2024, from 13% to 26% (Sidoti et al., 2025), all the while, in another survey of 1045 teens and parents, they report that most teachers prohibit or do not discuss these tools (Madden et al., 2024). This convergence of declining educational outcomes and rising reliance on AI tutoring tools highlights the need to assess the pedagogical advantages and limitations of AI-mediated instruction.

Advances in large language models (LLMs) have accelerated the development of intelligent tutoring systems that provide personalized feedback, generate explanations, and adapt to

learners' needs (Létourneau et al., 2025). Major AI and education organizations, such as OpenAI and Khan Academy, have introduced free tutoring systems designed to provide structured guidance, adapt to student input, and generate step-by-step reasoning (Khan, 2025; OpenAI, 2025). Controlled studies show that AI-generated feedback can yield learning outcomes comparable to human tutoring in mathematics (Pardos & Bhandari, 2024), while meta-analytic findings confirm that adaptive AI tutoring systems often outperform static digital instruction (Létourneau et al., 2025; Yi et al., 2024). These systems are increasingly accessible, scalable, and capable of producing domain-specific support at minimal marginal cost.

Personalization, the act of adapting aspects of the AI tutoring system to the user, is what makes AI tutoring responsive to the individual learner (Liu et al., 2025). Traditionally, personalization required complex programming to adjust difficulty, sequencing, or feedback timing to the individual learner (Huang et al., 2021). This led to improved engagement and comprehension by aligning content to the learner's current skill level (Yi et al., 2024). Modern systems allow this adaptation at the persona level through prompt engineering (Zheng et al., 2024). The tutor's communication style, tone, and instructional stance can be tailored to match individual preferences, creating a more responsive learning environment (Shumanov & Johnson, 2021; Zheng et al., 2024).

The potential to personalize the persona of an AI tutor, defined as its communicative style, tone, and structural approach to explanation, remains underexplored despite its relevance to learning efficiency. Studies in human-AI interactions show that personas change how learners interact with AI systems, yet effects on accuracy and cognitive effort vary and remain poorly understood

(Gkintoni et al., 2025; Zheng et al., 2024). Whether certain persona adaptations facilitate improved learning, deeper reasoning or reduce extraneous effort remains an open empirical question. The Felder-Silverman Learning Styles (FSLs) model provides a practical framework for investigating the outcomes of personalizing an AI tutor by adjusting its persona to match a user's preferred learning style (Felder, 1988; Graf et al., 2007). The potential to personalize an AI tutor's persona remains underexplored despite its relevance to learning efficiency. Although persona-level personalization is increasingly discussed, we still know little about how these adjustments affect cognitive effort (Gkintoni et al., 2025; Zheng et al., 2024). This limited understanding represents a first limitation in current research on AI tutoring.

Building on this point, a related limitation concerns the measurement of learning efficiency. Most evaluations of AI tutoring rely on accuracy scores, completion rates, or subjective assessments of mental effort. These measures provide limited insight into the cognitive processes that occur while learners engage with AI systems. From the perspective of Cognitive Load Theory (CLT), instructional formats can impose unnecessary processing demands, thereby increasing extraneous cognitive load and reducing the resources available for meaningful learning (Sweller, 2019; Sweller et al., 2011). If an AI tutor's communicative style does not align with a learner's processing preference, the learner must mentally reorganize explanations before integrating them with existing knowledge (Sweller, 2019). Dual-Process Theory (DPT) complements CLT by distinguishing between automatic, intuitive processing and deliberate, reflective reasoning (Evans & Stanovich, 2013). Early learning relies on reflective processes that place demands on working memory. When explanations are presented in a format consistent with a learner's preferred mode of processing, less reflective effort is required to interpret them

(Evans & Stanovich, 2013). Instruction that matches processing preference should thus support more efficient cognitive exertion on the assigned task. The cognitive cost of AI-mediated learning remains underexplored in educational research (Gkintoni et al., 2025). Continuous neurophysiological measures offer a promising opportunity to examine these cognitive processes with greater temporal precision (Antonenko et al., 2010).

Electroencephalography (EEG) and pupillometry provide complementary, high-resolution indices of moment-to-moment cognitive effort. Prior work identifies the frontal-parietal control network as central to sustained attention, working-memory manipulation, and executive control during learning (Cavanagh & Frank, 2014; Klimesch, 1999, 2012). Increases in frontal midline theta activity and decreases in parietal alpha activity track heightened task demands, and the ratio of these oscillatory components has been proposed as a robust indicator of task load in human-computer interaction (Sollazzo et al., 2025). Pupillometry offers an additional index of arousal and cognitive effort, with dilation patterns sensitive to processing load and attention (Krejtz et al., 2018; Rozado & Dunser, 2015). Despite their value, these measures are rarely employed in studies of AI tutoring, leaving a gap in understanding the cognitive dynamics that accompany AI-mediated instruction (Gkintoni et al., 2025).

To test the effectiveness of persona-level adaptations in AI tutors, we've chosen to use the Felder-Silverman Learning Styles (FSLs) model as a conceptual framework. Although its psychometric properties are debated, the model formalizes differences in how learners prefer to perceive and organize information (Felder & Je, 2005; Graf et al., 2007; Haug et al., 2023). If persona-level adjustments that align with FSLs profiles improve cognitive efficiency, this would

indicate that the value lies less in the model itself and more in the way instructional alignment influences mental effort. This possibility raises important questions about the principles guiding personalization in AI tutors and the role of surface-level adaptation in reducing extraneous load.

This thesis addresses these gaps by examining the neurophysiological and behavioural effects of persona-level personalization in an AI mathematics tutor. Using a mixed-design experiment, the study investigates whether adapting the tutor's communicative style to match learners' FSLs profiles alters cognitive effort, engagement, and learning performance. The experiment integrates EEG and pupillometry with behavioural indicators such as non-response rates and formula use, as well as subjective assessments of task load and satisfaction. By combining these measures, the study evaluates not only what learners achieve but also how efficiently they process information during AI tutoring.

The primary research question guiding this work is:

To what extent does personalized AI-mediated tutoring, adapted to learners' Felder-Silverman profiles, impact the user's cognitive, behavioural, and perceived experience?

To investigate this question, we conducted a laboratory experiment in which participants completed three blocks of mathematics tutoring with either a personalized or a non-personalized AI tutor. Personalization was implemented through system-level prompt engineering that adjusted explanatory style, according to FSLs categories. To examine the effects of these mechanisms, a custom AI tutor interface was made to enable precise manipulation of persona-level prompts, consistent delivery of condition-specific instructions, and automatic logging of all conversational exchanges for later prompt-level analysis. Cognitive load was assessed using

frontal-theta/parietal-alpha EEG ratios derived from sixteen electrode combinations, as well as pupillometric measures of task-evoked dilation (Sollazzo et al., 2025). Behavioural engagement was quantified through response patterns during tutoring, and learning outcomes were measured through post-task accuracy tests. Together, these data provided a multimodal profile of users' cognitive and behavioural responses to personalization.

Our results show that while persona-level personalization did not significantly improve accuracy or satisfaction, it produced measurable reductions in cognitive effort that persisted beyond the LLM interactions. Participants interacting with the personalized tutor exhibited lower EEG task load ratios and reduced pupil dilation during instruction. They also exhibited greater behavioural elaboration, as evidenced by more frequent use of mathematical notation and fewer unanswered prompts. These patterns suggest that personalization may support more efficient allocation of cognitive resources, even when immediate learning gains are not apparent. This distinction between performance and cognitive efficiency aligns with DLP and CLT's prediction that well-aligned instruction reduces extraneous load and demands on Type 2 processes, allowing learners to direct effort toward schema formation and, over time, engage Type 1 processes in a more efficient manner (Evans & Stanovich, 2013; Sweller, 2019).

This thesis contributes to ongoing research on AI-mediated learning in three ways. First, it provides empirical evidence that persona-level adaptation influences neurophysiological indicators of learning efficiency, suggesting that surface-level communicative adjustments can shape users' cognitive states. Second, it demonstrates the feasibility of integrating continuous EEG and pupillometry into controlled studies of AI tutoring, offering methodological guidance

for future work on cognitive processes during human-AI interaction. Third, it highlights the distinction between traditional learning measures and cognitive efficiency, emphasizing the need for multidimensional evaluation frameworks when assessing adaptive AI systems.

The findings also have application and policy relevance. Private tutoring is a growing global market, valued at roughly 92 billion in 2022 and projected to reach 155 billion by 2030 (Grand View, 2025). These figures highlight unmet demand for affordable, individualized learning support. If public education systems can integrate safe and pedagogically aligned AI-mediated learning tools, they can improve access and equity without increasing costs.

Schools are already moving from pilot projects to permanent deployment. Khan Academy launched its research-informed personalized AI tutor, Khanmigo, which now supports more than two million active users and is expanding to more than 380 school districts (Khan, 2025). As more districts embrace AI-mediated learning, early design decisions will shape long-term practice and learning outcomes. Research on cognitive efficiency can inform the development of AI tutors that enhance learning while preserving the effort that sustains it. Evidence from this study can guide policymakers, educators, and designers toward AI tutors that support effective and efficient learning.

Thesis Structure

The thesis is organized into three main chapters, followed by a bibliography and appendices. Chapter 1 introduces the research context, explaining the motivation, theoretical background, and the role of AI-mediated learning in the form of AI-tutors addressing educational challenges.

Chapter 2 presents the empirical article, which examines the results of an experiment implementing persona-level personalization of an Intelligent tutoring system in a math learning context. Chapter 3 synthesizes the findings, outlines theoretical and practical implications, and discusses limitations and directions for future research. The bibliography lists all cited works for each chapter, and the appendices include supplementary materials relevant to data collection and analysis.

Personal Contribution

Table 1 summarizes my own intellectual contribution to each component of the research project and thesis writing. Under the lab’s guidelines, students are expected to contribute at least 50% of the overall work. Any area in which my contribution is greater than 50% indicates that I took primary responsibility and demonstrated leadership in that phase.

Table 1. *Personal Contributions and Responsibilities Throughout the Research Project*

Research Activity	Contribution
Research Questions	Identifying the research problem and formulating research questions aligned with AI-mediated learning and cognitive-load measurement – 60% • Initial problem framing and contextualization • Research questions refined with supervisors
Literature Review	Conducting the full systematic literature review – 100% • Search strategy, screening, synthesis, and drafting of all sections on AI tutors, CLT, FSLs, EEG, and dual-process theory

Experimental Design	Designing and formalizing the experimental protocol – 50% • Received initial direction and collaboratively came to designing the mixed $2 \times 2 \times 3$ design • Members of the lab provided a template for the protocol, from which we adapted to our project
Stimuli / Tutoring Materials	Creating the AI tutoring persona level prompts and math questions – 90% • Core and personalized system prompts • Topic-specific tutoring material for Venn diagrams, logarithms, and parabolas • Pre-testing adjustments to ensure stable LLM behaviour Nback Paradigms preparation 15% - Provided by the lab
Questionnaires	Building all questionnaires in Qualtrics – 95% • Finding condensing and organizing perceptual questionnaire items in the experimental design. • Pre/post math assessments Received support to code an in-house tool for FSLs response calculation– 10%
Ethics	Preparing and submitting ethics applications – 10%
Pre-tests	Running internal pre-tests – 85% • Pilot testing of prompts, EEG setup, task timing, and interface usability. Received guidance from the Tech3lab team on how to use lab tools and problem-solving. • Minor adjustments to improve data quality and session flow
Recruitment	Recruiting and managing participants – 60% • Inclusion/exclusion criteria, registration form, and panel campaign management.

	• Participant communication and scheduling (supported by lab operations)
Data Collection	Full responsibility for data collection – 50% • Moderation of half of the sessions, with scheduling support from Tech3Lab research assistants • EEG setup, signal-quality checks, pupil calibration • Management of AI interface, timing, and contingency handling • Cleaning and resetting equipment and rooms
Data Analysis	Cleaning and analyzing all behavioural, questionnaire, and physiological data – 60 % • Statistical modelling and results verification with support from the lab statistician • EEG preprocessing and theta-alpha ratio computation were done by Tech3Lab Post-doc – 10%
Writing	Writing thesis and article-style manuscript – 100% • Drafting all sections and integrating co-author revisions

References

- Antonenko, P. P., Paas, F., Grabner, R., & Gog, T. (2010). Using Electroencephalography to Measure Cognitive Load. *Educational Psychology Review*, 22, 425-438. <https://doi.org/10.1007/s10648-010-9130-y>
- Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, 18(8), 414-421. <https://doi.org/10.1016/j.tics.2014.04.012>
- Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How People Use ChatGPT. *National Bureau of Economic Research Working Paper Series*, No. 34255. <https://doi.org/10.3386/w34255>
- Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3), 223-241. <https://doi.org/10.1177/1745691612460685>
- Felder, R. (1988). Learning and Teaching Styles in Engineering Education. *Journal of Engineering Education*, 78, 674-681.
- Felder, R., & Je, S. (2005). Applications, reliability and validity of the Index of Learning Styles. *International Journal of Engineering Education*, 21, 103-112.
- Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy. *Brain Sci*, 15(2). <https://doi.org/10.3390/brainsci15020203>
- Graf, S., Viola, S. R., Leo, T., & Kinshuk. (2007). In-Depth Analysis of the Felder-Silverman Learning Style Dimensions. *Journal of Research on Technology in Education*, 40(1), 79-93. <https://doi.org/10.1080/15391523.2007.10782498>
- Grand View Research Group. (2025). *Private Tutoring Market Size, Share & Growth Report (2023-2030)*. <https://www.grandviewresearch.com/industry-analysis/private-tutoring-market-report>
- Haug, J., Fischer, D., & Hagel, G. (2023). Development of a short form of the index of learning styles for the use in adaptive learning systems. Proceedings of the 5th European Conference on Software Engineering Education.
- Huang, Y., Lobczowski, N. G., Richey, J. E., McLaughlin, E. A., Asher, M. W., Harackiewicz, J. M., Aleven, V., & Koedinger, K. R. (2021). *A General Multi-method Approach to Data-Driven Redesign of Tutoring Systems* LAK21: 11th International Learning Analytics and Knowledge Conference, Irvine, CA, USA. <https://doi.org/10.1145/3448139.3448155>, <https://dl.acm.org/doi/pdf/10.1145/3448139.3448155>

- Khan Academy. (2025). *Annual Report SY 24-25: July 1 2024–June 30 2025*.
<https://annualreport.khanacademy.org>
- Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews*, 29(2), 169-195.
[https://doi.org/10.1016/S0165-0173\(98\)00056-3](https://doi.org/10.1016/S0165-0173(98)00056-3)
- Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 16(12), 606-617.
<https://doi.org/10.1016/j.tics.2012.10.007>
- Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE*, 13(9), e0203629. <https://doi.org/10.1371/journal.pone.0203629>
- Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10(1), 29. <https://doi.org/10.1038/s41539-025-00320-7>
- Liu, B., Zhang, J., Lin, F., Jia, X., & Peng, M. (2025). One size doesn't fit all: A personalized conversational tutoring agent for mathematics instruction. Companion Proceedings of the ACM on Web Conference 2025,
- Madden, M., Calvin, A., Hasse, A., & Lenhart, A. (2024). *The dawn of the AI era: Teens, parents, and the adoption of generative AI at home and school*.
https://www.common sense media.org/sites/default/files/research/report/2024-the-dawn-of-the-ai-era_final-release-for-web.pdf
- OECD. (2023a). *PISA 2022 results (Volume I and II) - Country notes: Canada*. OECD Publishing. Retrieved September, 2025 from https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/canada_901942bb-en.html
- OECD. (2023b). *PISA 2022 results (Volume I and II) - Country notes: United States*. OECD Publishing. Retrieved Sept 2025 from https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/united-states_a78ba65a-en.html
- OpenAI. (2025, July 29, 2025). *Introducing study mode*. openAI. Retrieved Sept, 2025 from <https://openai.com/index/chatgpt-study-mode>
- Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, 19(5), e0304013. <https://doi.org/10.1371/journal.pone.0304013>

- Rozado, D., & Dunser, A. (2015). Combining EEG with Pupillometry to Improve Cognitive Workload Detection. *Computer*, 48(10). <https://doi.org/10.1109/MC.2015.314>
- Shumanov, M., & Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in Human Behavior*, 117, 106627.
- Sidoti, O., Park, E., & Gottfried, J. (2025). About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023. *Pew Research Center Short Reads*. <https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/>
- Sollazzo, K., Karran, A. J., Rolon-Merette, T., Stefanescu, I. M., Coursaris, C., Léger, P.-M., & Sénécal, S. (2025, 2025/12). *Pathways to Optimal Learning: Task Load-Driven Neuroadaptive Adaptation and Motivational Incentives* NeuroIS Retreat 2025,
- Sweller, J. (2019). Cognitive load theory and educational technology. *Educational Technology Research and Development*, 68(1), 1-16. <https://doi.org/10.1007/s11423-019-09701-3>
- Sweller, J., Ayres, P. L., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- UNESCO. (2023). *Highlights from the global report on teachers: What you need to know*. <https://www.unesco.org/en/articles/highlights-global-report-teachers-what-you-need-know>
- Yi, L., Liu, D., Jiang, T., & Xian, Y. (2024). The Effectiveness of AI on K-12 Students' Mathematics Learning: A Systematic Review and Meta-Analysis. *International Journal of Science and Mathematics Education*, 1-22. <https://doi.org/10.1007/s10763-024-10499-7>
- Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2024). When” a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large language models. Findings of the Association for Computational Linguistics: EMNLP 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.888>

Chapter 2: Neurophysiological effects of AI-mediated mathematics education: Exploring EEG measures of task load and the user's experience

Uriel Karerwa, Alexander John Karran, Sylvain Sénécal, Thaddé Rolon-Mérette

Abstract

Artificial intelligence-mediated learning has gained prominence as learners increasingly rely on large language models. However, little is known about the neurophysiological mechanisms through which personalization influences cognitive load during AI-guided instruction. In this study, personalization refers to adapting the tutor's communicative style, tone, and explanatory structure to align with learners' profiles as defined by the Felder-Silverman Learning Style Model. This study examined whether adapting the persona of a large language model tutor to match learners' profiles from the Felder-Silverman Learning Style model affects learning outcomes, behavioural engagement, and task load. Forty participants completed three blocks of mathematics tutoring under treatment and control conditions, with electroencephalography (EEG), pupillometry, and psychometric measures being recorded. Task load was operationalized as the ratio of frontal theta (4-7 Hz) to parietal alpha (8-13 Hz) power across sixteen electrode pairs, complemented by changes in pupil dilation and NASA Task Load Index ratings. The results demonstrated that persona-level personalization did not significantly improve learning accuracy ($p = 0.42$) or satisfaction ($p = 0.315$). However, personalized interactions produced lower frontal theta to parietal alpha ratios ($p < 0.001$) and reduced pupil dilation ($p = 0.003$), indicating improved cognitive efficiency during instruction. Behavioural indicators, including

reduced nonresponse rates ($p = 0.095$) and increased use of mathematical notation ($p = 0.059$), further suggested deeper engagement. These findings indicate that personalization influences the efficiency of cognitive resource allocation even when immediate performance remains unchanged. The results highlight the value of integrating neurophysiological measures into the study of adaptive artificial intelligence tutors and suggest that persona-level adaptations shape the cognitive demands of human AI learning interactions.

1. Introduction

Researchers are characterizing the global decline in math performance as a "learning crisis" with profound implications for social mobility, economic development, and workforce readiness in science, technology, engineering, and mathematics (STEM) fields (Létourneau et al., 2025; Spitzer & Moeller, 2024). Global student performance in mathematics has declined precipitously despite record investment in education systems (OECD, 2023a, 2023b; UNESCO, 2023). The Program for International Student Assessment (PISA) reveals that between 2002 and 2022, mathematics scores among North American students fell to their lowest recorded levels, with Canadian and American students consistently underperforming baseline proficiency thresholds (OECD, 2023a, 2023b).

Artificial intelligence-mediated learning has become widespread and is increasingly recognized as an effective instructional approach. By 2025, hundreds of millions of people interact with large language models (LLMs) every day, with approximately one-third of all use cases related to a learning task (Chatterji et al., 2025). Advances in large language models (LLMs) have enabled a new generation of AI-mediated learning systems, such as intelligent tutoring systems (ITS), that deliver personalized, conversational feedback. In July 2025, OpenAI, the creators of ChatGPT, released an ITS that scaffolds learning through strategies such as Socratic questioning and stepwise explanations (OpenAI, 2025). Meta-analyses show that outcomes achieved with AI-mediated tutoring are comparable to or exceed those of many traditional educational interventions, underscoring its potential as a scalable complement to classroom instruction

(Létourneau et al., 2025; Yi et al., 2024). Together, these findings position AI-mediated tutoring as a scalable instructional approach with growing relevance for education systems in crisis.

Growing evidence indicates that personalizing AI-mediated learning, via AI-enabled ITS can further enhance their instructional effectiveness. Compared with non-intelligent learning platforms, ITSs that incorporate adaptive features consistently yield stronger learning outcomes (Létourneau et al., 2025). Personalization mechanisms that adjust content to learners' prior knowledge, pacing, or topical interests have been shown to sustain motivation, increase engagement, and deepen comprehension (Létourneau et al., 2025; Yi et al., 2024). Despite these advances, personalization remains conceptually narrow. The potential of persona-level adaptation, defined as the dynamic adjustment of a tutor's expressed identity, tone, and communicative style to align with user-provided parameters such as their preferred interactional and cognitive style, has received little empirical investigation, particularly concerning the neurophysiological mechanisms that drive its effects on cognitive load and learning dynamics (Gkintoni et al., 2025; Liu et al., 2025; Zheng et al., 2024).

Three critical gaps limit current understanding of AI-mediated learning. First, most AI tutoring research relies exclusively on task outcomes or self-reports, which provide little insight into the cognitive processes underlying performance (Gkintoni et al., 2025; Sarailoo et al., 2022). A tutor that improves accuracy but taxes working memory may appear effective in the moment yet impede long-term learning (Gkintoni et al., 2025; Sweller, 2019). Second, persona-level personalization (in which the AI tutor's output style is adapted using a system-level prompt that alters the LLM's objective function) has not been tested using validated learning-style

frameworks and neurophysiological measures. Third, Cognitive Load Theory (CLT) predicts that personalized AI-mediated learning should reduce extraneous cognitive load and optimize germane cognitive load (Sweller, 2019; Sweller et al., 2011). However, this hypothesis has not been tested with converging, continuous neurophysiological indices such as electroencephalography (EEG) and pupillometry.

One promising avenue for advancing AI-mediated tutoring is to improve how these systems personalize instruction to individual learners through persona design. One potential method of providing such persona personalization is through established learning-style frameworks. The Felder-Silverman Learning Styles (FSLs) model provides a tractable framework for testing persona-level personalization. Although its psychometric validity has been questioned (Felder & Je, 2005; Graf et al., 2007; Haug et al., 2023) its continued application provides a useful boundary condition: if personalization grounded in a contested model yields measurable cognitive or learning benefits, this would suggest that the value of personalization lies less in the model itself and more in its capacity to influence key dimensions of human-AI interaction, such as cognitive load.

Personalization was implemented through system-level prompting, which refers to a set of instructions that defines the global behaviour, tone, and communicative style of the large language model across an entire interaction. These instructions operate at a level above user-level inputs, shaping how the model interprets and responds to subsequent prompts (Zheng et al., 2024). To operationalize this mechanism, we used an existing large language model and added a custom AI tutor interface that allowed controlled modification of persona-related system

prompts, ensured consistent delivery of the required instructional conditions, and captured all conversational exchanges for subsequent prompt-level analysis.

The goal of this study is to investigate whether persona-level adaptation of an AI mathematics tutor to learners' Felder-Silverman profiles enhances learning effectiveness, reduces cognitive load, and improves user experience. We thus motivate this study by posing the following research question: To what extent does personalized AI-mediated tutoring, adapted to learners' Felder-Silverman profiles, affect users' cognitive, behavioural, and perceived experiences?

We operationalize task load as one's cognitive load while completing a task, measured using continuous EEG and pupillometric indices during tutoring interactions, alongside self-reported mental effort and learning outcomes. By combining neurophysiological measures with behavioural and perceptual data, this study provides a comprehensive assessment of how personalization shapes both what students learn and how efficiently they learn it.

The results of this study indicate that the AI tutor personalized to learners' Felder-Silverman profiles did not significantly improve learning accuracy or satisfaction. However, it yielded measurable gains in cognitive efficiency, as reflected by lower physiological task-load indices from EEG and pupillometry. Behaviourally, participants in the personalized learning condition showed greater interactional elaboration, evidenced by lower non-response rates and a greater proportion of prompts containing mathematical formulas. These results suggest that personalization optimizes cognitive resource allocation during learning, facilitating more efficient processing and deeper engagement with content, even without immediate improvements in perceived experience or performance.

2. Related Work

This section situates the study within prior research on AI-mediated instruction and theoretical and cognitive models of learning efficiency. It first reviews AI-mediated learning and intelligent tutoring systems to clarify how contemporary LLM-enabled tutors relate to established adaptive frameworks. It then outlines the theoretical foundations that inform predictions about cognitive efficiency during tutoring interactions, focusing on Cognitive Load Theory and Dual-Process Theory. Subsequently, it examines neurophysiological and ocular indicators relevant to task load in human-computer learning contexts, followed by an analysis of the Felder-Silverman instructional design model. The section concludes with a review of learning and user-experience metrics used to evaluate instructional effectiveness. Together, these strands provide the conceptual and empirical rationale for examining persona-level adaptation and its effects on cognitive effort, engagement, and performance.

2.1 AI-Mediated Learning and Intelligent Tutoring Systems (ITS)

AI-mediated learning refers to instructional processes in which neural network models, such as large language models, structure, deliver, or adjust learning activities through natural language. These systems analyze learner input during conversation and adapt their responses in real time. Although reviews often separate intelligent tutoring systems (ITS), adaptive learning systems (ALS), pedagogical agents, and conversational LLM tutors, the boundary between ITSs and LLM-based tutors has largely dissolved (Létourneau et al., 2025; Yi et al., 2024). Historically, ITSs relied on rule-based or model-based engines to track learner knowledge and select appropriate responses (Walkington & Bernacki, 2019). Today, many of these functions are

implemented directly through conversational interaction with LLMs. An ITS can therefore be considered a conversational LLM tutor when the model provides feedback, worked examples, hints, or dialog-based support (Yi et al., 2024). Contemporary reviews treat LLM tutors as an evolution or subclass of ITSs because both pursue the same instructional goal: diagnosing understanding and adapting instruction (Létourneau et al., 2025; Yi et al., 2024). Consequently, the present thesis uses the terms ITS and AI tutor to refer to the same class of systems that enable AI-mediated learning.

Across these systems, the effectiveness of AI-mediated learning depends on how it is adapted to the learner. Meta-analytical reviews show that the strongest learning gains emerge when systems use learner data to adjust difficulty, pacing, and support in real time, as in the case of ALS (Yi et al., 2024). Létourneau et al. (2025) report that ITSs outperform non-adaptive AI-learning tools when they include features such as targeted hints, real-time error detection, and personalized practice schedules (Létourneau et al., 2025). Likewise, studies of ChatGPT-generated help demonstrate that when hallucination-reduction techniques and quality controls are applied, LLM-produced hints can achieve learning gains comparable to human-written help (Pardos & Bhandari, 2024).

Because LLM-driven tutors do not rely on fixed scripts, their instructional behaviour is shaped primarily through system-level prompts that define the tutor's role, tone, and reasoning style. A prompt is any input text provided to a model to guide its output, and prompting is the process of crafting such inputs to influence how the model responds in an instructional setting. Prompting influences how the model interprets questions, structures explanations, and determines the level

of support to provide (Zheng et al., 2024). Findings from persona-prompt research show that while personas can change interactional style, how a tutor “sounds”, they do not reliably improve accuracy and may introduce unpredictable variation across different types of questions (Zheng et al., 2024). The same study shows that domain-aligned personas and other persona distinctions yield small effects, but that there is no single reliable method for automatically selecting beneficial personas. We can thus infer from Zheng et al. (2024) that a few practices, namely, implementing domain-specific persona instructions grounded in pedagogical evidence and repeatedly tested, are more likely to yield better outcomes.

The literature suggests that adaptivity and personalization can operate at multiple levels: content selection, feedback timing, explanation scaffolding, and discourse structure (Létourneau et al., 2025; Yi et al., 2024). Persona-level adaptation occupies the least explored of these layers: the communicative framing of instruction, including tone, style, and relational cues (Liu et al., 2025; Zheng et al., 2024).

Emerging evidence indicates that AI-tutoring systems that adjust the conversational style of the persona using frameworks such as the Felder-Silverman model increase engagement (Liu et al., 2025). This body of research provides a foundation for studying whether system-prompt-driven personas can modulate task load, facilitate more fluent learner-tutor interactions, and ultimately support more efficient learning, which motivates the present investigation.

While AI-mediated tutoring can improve learning under certain conditions, its effects depend on aspects of the user's experience, like how they process instructional information during the

interaction. This consideration motivates the learning-efficiency theories that specify mechanisms linking instructional format to cognitive effort.

2.2 Theoretical Framework: Cognitive Load Theory

Cognitive Load Theory (CLT) provides a framework for predicting how instructional design influences learning efficiency (Sweller, 2019; Sweller et al., 2011). CLT distinguishes three forms of cognitive load: intrinsic, extraneous, and germane. Intrinsic load reflects the inherent complexity of the material. Extraneous load reflects unnecessary cognitive effort imposed by suboptimal instructional design. Germane load reflects effort devoted to schema construction and automation. Effective instructional design seeks to minimize extraneous load and increase germane processing, thereby allocating limited working memory resources to meaningful processing (Sweller, 2019).

In AI-mediated learning through a chat interface, personalization strategies can plausibly reduce extraneous load by aligning instructional delivery with learners' processing preferences (Sweller, 2019). For example, a learner who prefers sequential explanations may experience increased extraneous load when presented with global, connection-oriented explanations, as they must expend additional cognitive effort to translate the information into their preferred format (Graf et al., 2007; Sweller, 2019). Conversely, when the explanation structure better matches the learner's preferences, less effort is required to interpret the instruction, which should improve learning efficiency under CLT.

CLT suggests that persona-level personalization could lower extraneous cognitive load and enhance learning efficiency. Nevertheless, this hypothesis and the application of the FSL model have not been validated through continuous neurophysiological measurements during AI tutoring sessions. Most CLT studies rely on subjective assessments, such as the NASA Task Load Index (Hart & Staveland, 1988), or on behavioural indicators, such as secondary-task performance. Although useful, these measures have limited temporal resolution and may not detect short-term variations in cognitive load experienced during learning.

To maintain terminological consistency, references to cognitive load will be used only when discussing prior work that addresses Cognitive Load Theory directly, whereas task load will be used throughout this study to refer to the specific construct which was measured and analyzed. Task load refers exclusively to the momentary cognitive demands imposed by the learning activity and is further defined in the methods section.

Building on these theoretical distinctions, the next section defines the specific form of cognitive demand examined in this study and clarifies its relationship to the broader construct of cognitive load.

2.3 Theoretical Framework: Dual-Process Theory

Complementary to CLT in describing the mental efficiency of AI-mediated learning, Dual-Process Theory (DPT) distinguishes between two modes of cognition, Type 1 and Type 2, that jointly shape learning and reasoning (Evans & Stanovich, 2013). Type 1 processes are fast, automatic, and independent of working memory, generating intuitive responses from prior

experience. Type 2 processes are slower and deliberate, relying on working memory to support cognitive decoupling, or reasoning beyond immediate perception (Evans & Stanovich, 2013).

Learning depends on the interaction of these systems. Early in learning, Type 2 reasoning dominates as learners analyze and build schemas; with practice, knowledge becomes automated under Type 1, enabling quick, low-effort responses. This reflects Evans and Stanovich's default-interventionist model, where Type 1 guides routine actions and Type 2 intervenes during novelty or conflict (Evans & Stanovich, 2013). Cognitive Load Theory (CLT) complements DPT by describing how limited working memory capacity constrains learning efficiency.

Extraneous load entails unnecessary Type 2 activation of information that is better handled automatically. Germane load reflects productive Type 2 use of the schema for construction. Automation, central to both theories, marks the shift from controlled (Type 2) to automatic (Type 1) processing. When an AI tutor adapts its communication style to the learner's processing preference, the learner expends less reflective effort to interpret information. This efficiency is reflected in lower cognitive-load indices, indicating that mental resources are used more effectively. Effective engagement of Type 2 reasoning occurs when working memory is not overloaded, allowing effortful processing to contribute to schema construction. As schemas consolidate, performance shifts toward Type 1 automaticity, reducing cognitive effort, but only after repeated practice.

In essence, Dual-Process Theory describes how learners move from reflective to automatic processing, whereas Cognitive Load Theory explains how limited working memory resources are distributed during this shift. Together, they clarify how personalization can lower cognitive load.

Theories of learning efficiency suggest that personalization influences cognitive effort during interaction. Therefore, the next subsection reviews neurophysiological and ocular indicators used to capture task-related cognitive effort during human-AI learning.

2.4 Neurophysiological Indicators of Task Load

Determining which neural systems best reflect fluctuations in cognitive effort during learner-AI interactions is a central challenge in educational neuroscience (Antonenko et al., 2010; Gkintoni et al., 2025; Karmakar et al., 2023). Effective monitoring requires indices that are sensitive to dynamic shifts in attention, working memory, and executive control as learners engage with adaptive tutoring systems. Converging evidence from neuroimaging and electrophysiological studies identifies the frontal-parietal control network as the primary substrate supporting these operations (Cavanagh & Frank, 2014; Klimesch, 1999). This distributed network coordinate's goal maintenance, top-down attention, and information integration during effortful learning (Cavanagh & Frank, 2014; Wascher et al., 2014). Its dual involvement in maintaining task rules and integrating sensory information makes it an appropriate target for neurophysiological assessment of task load in human-AI learning contexts.

Within this system, frontal midline theta oscillations (4-7 Hz) consistently increase with cognitive control demands (Cavanagh & Frank, 2014; Klimesch, 1999; Wascher et al., 2014). Activity in the medial and dorsolateral frontal cortices rises when learners must sustain attention, manipulate information in working memory, or inhibit irrelevant thoughts (Cavanagh & Frank, 2014; Corbetta & Shulman, 2002). Frontal theta is therefore interpreted as a marker of active top-down regulation: it increases when the learner must, for example, hold the tutor's

instructions in working memory while simultaneously reasoning through a problem. This sensitivity to controlled processing makes frontal theta a reliable indicator of cognitive effort in complex, interactive learning environments.

Complementing this pattern, posterior parietal alpha activity (8-13 Hz) reflects the level of attentional engagement with sensory and visuospatial information (Capotosto et al., 2009; Klimesch, 2012). Decreases in parietal alpha power (desynchronization) accompany heightened visual attention and information integration, while relative alpha synchronization is observed during fatigue, mind-wandering, or disengagement (Capotosto et al., 2009; Klimesch, 2012). The parietal cortex, particularly the inferior and superior parietal lobules, thus provides a functional counterpoint to frontal control regions by indexing the extent to which sensory and spatial resources are recruited during learning.

Given these complementary dynamics, a growing number of studies use a ratio of frontal theta to parietal alpha power, capturing the balance between executive control and attentional engagement (Sollazzo et al., 2025). Elevated frontal-theta/parietal-alpha ratios have been associated with increased task load across diverse tasks, including multimedia learning and human-computer interaction (Sollazzo et al., 2025). This ratio-based approach provides a temporally sensitive index of cognitive efficiency that reflects the interplay between sustained effort and perceptual processing, making it particularly suitable for evaluating task load during adaptive AI-tutoring sessions.

Electroencephalography (EEG) is well suited to measuring task load due to its millisecond-level temporal resolution and sensitivity to oscillatory dynamics in the theta and alpha frequency

bands (Antonenko et al., 2010). Frontal theta power has been consistently associated with sustained attention and working memory demands, while parietal alpha power decreases with fatigue and disengagement (Antonenko et al., 2010; Klimesch, 1999). The resulting frontal-theta to parietal-alpha ratio thus serves as a sensitive indicator of moment-to-moment fluctuations in task load.

To diminish the likelihood of capturing noise at individual electrode sites and increase our chances of capturing the distributed activation of fronto-parietal dynamics Sollazzo (2025) proposes computing 16 theta-alpha ratios across all combinations of four frontal electrodes (Fp1, Fp2, F3, F4) and four parietal electrodes (CP1, CP5, P3, Pz). The ratios are averaged across all pairs, as opposed to compute theta-alpha ratios from single electrode pairs.

Pupillometry provides a complementary physiological index of task load. Changes in pupil diameter have been reliably linked to cognitive effort, with dilation reflecting increased task demands (Krejtz et al., 2018). The pupillary response is mediated by the locus coeruleus-norepinephrine system, which modulates arousal and attentional (Rozado & Dunser, 2015). Together, EEG-derived indices and pupillometry provide temporally sensitive measures that can complement self-report workload measures.

Neurophysiological indicators can quantify the cognitive demand of an interaction. However, the design rationale for persona-level adaptation requires a framework that specifies how communication styles differ across learners. The next subsection motivates the Felder-Silverman framework as the basis for persona adaptations in this thesis.

2.5 Felder-Silverman Learning Style Model

The Felder-Silverman Learning Style (FSLs) model was developed within engineering education to account for systematic differences in how students perceive and organize information (Felder, 1988). The model integrates elements from cognitive psychology and instructional theory, including information-processing accounts of learning, experiential learning theory, and Jungian psychological type theory. It conceptualizes learning styles as relatively stable preferences along bipolar dimensions that shape how learners engage with instructional material.

The FSLs model emerged in response to observed mismatches between conventional engineering instruction and diverse student preferences. Traditional engineering pedagogy emphasized deductive presentation and verbal exposition. Felder and Silverman argued that such practices privileged certain learners while disadvantaging others whose preferences aligned with concrete examples, visual representations, or incremental structure (Felder, 1988). Thus, the model was designed primarily as a pedagogical heuristic to inform instructional design and improve inclusivity in technical education.

The theoretical foundations of FSLs draw from multiple strands of educational theory. First, the Active-Reflective dimension reflects Kolb's experiential learning cycle, which distinguishes between active experimentation and reflective observation (Kolb, 1984). Second, the Sensing-Intuitive and Sequential-Global dimensions incorporate concepts from cognitive style research and Jungian typology (Felder, 1988; Graf et al., 2008). Third, the Visual-Verbal dimension aligns with dual coding theory and multimedia learning research, which emphasize the role of representational formats in cognitive processing (Felder, 1988; Paivio, 1986). Consequently,

FSLs represents a hybrid model that connects learner preferences to established theories of cognition and instruction.

The model operationalizes four dimensions: Active-Reflective, Sensing-Intuitive, Visual-Verbal, and Sequential-Global. The Index of Learning Styles (ILS) was later developed to measure these dimensions and has been examined for reliability and construct validity across disciplines (Felder & Spurlin, 2005; Graf et al., 2007). Empirical evaluations report moderate internal consistency and a factorial structure consistent with the proposed dimensions, although effect sizes vary across contexts. For instance, Graf et al. (2007) conducted an in-depth analysis of the FSLs dimensions in technology-enhanced learning environments and found partial support for dimensional independence and predictive utility in adaptive systems.

Despite its pedagogical influence, the broader learning styles literature remains contested. Reviews in high-impact journals have questioned the empirical basis of the “meshing” hypothesis, which posits that matching instructional mode to learner style improves achievement (Pashler et al., 2008). These critiques emphasize limited experimental evidence demonstrating crossover interactions between style and instructional method. Therefore, contemporary scholarship increasingly distinguishes between describing learner preferences and asserting causal effects on learning outcomes.

Within this debate, the FSLs model occupies a pragmatic position. Rather than claiming that style matching directly enhances achievement, several studies examine its use in adaptive educational systems to personalize presentation structure and interaction patterns (Graf et al., 2007; Haug et al., 2023). Recent work in intelligent tutoring and AI-supported learning

environments situates FSLs-based adaptation within broader frameworks of learner modelling and personalization (Zhang et al., 2025). In this context, style dimensions function as input variables for adaptive scaffolding rather than deterministic predictors of performance.

The present thesis adopts this restrained interpretation. FSLs profiles are treated as indicators of communication preference that may influence perceived cognitive effort during AI-mediated instruction. No claim is made that the model captures stable cognitive capacity or that instructional matching ensures superior achievement. Instead, the experimental manipulation examines whether alignment between tutor communication and selected FSLs dimensions affects cognitive efficiency, operationalized as reduced mental effort for comparable performance.

The manipulation targets two dimensions that are theoretically relevant to instructional structure: Sensing-Intuitive and Sequential-Global. The Sensing-Intuitive dimension differentiates preference for concrete, example-driven explanations versus abstract, principle-oriented framing. This distinction corresponds to differences in schema formation and generalization processes described in cognitive psychology. The Sequential-Global dimension differentiates incremental, stepwise organization from a holistic overview-oriented presentation. This distinction aligns with analytic versus holistic processing styles and with research on advanced organizers and scaffolding in instructional design.

The remaining dimensions, Active-Reflective and Visual-Verbal, are described for completeness but are not central to the experimental manipulation. The Active-Reflective dimension concerns engagement through discussion and action versus internal deliberation. The Visual-Verbal

dimension concerns representational modality, which is less directly modifiable through text-based tutor prompts in the present design.

By situating the FSLS model within established theories of cognition, acknowledging critiques of the learning styles literature, and limiting interpretive claims to communication alignment and cognitive efficiency, this section clarifies the theoretical and empirical basis for its inclusion. This framing ensures conceptual rigour while preserving the practical rationale for testing style-informed adaptation in AI-mediated tutoring.

2.6 Learning Metrics and User Experience

Learning outcomes in computer-based environments can be conceptualized along two complementary dimensions: cognitive outputs (what learners know or can do) and behavioural processes (how learners engage with learning opportunities). Traditional assessments emphasize cognitive products, including accuracy, correct response rates, and knowledge retention, which reflect the accumulation of declarative and procedural knowledge (Alrashidi et al., 2016; Anderson, 1982). However, these product measures provide only partial insight into the learning experience, as they do not capture the behavioural and motivational processes that mediate knowledge acquisition. Behavioural indicators such as completion rates (the proportion of assessment items attempted), nonresponse rates (the inverse of completion, including both blank responses and explicit "I don't know" selections), and formula use during tutoring interactions represent process measures that illuminate learners' engagement, self-regulation, and cognitive involvement with instructional materials (Alrashidi et al., 2016; Blumenfeld & Paris, 2004)

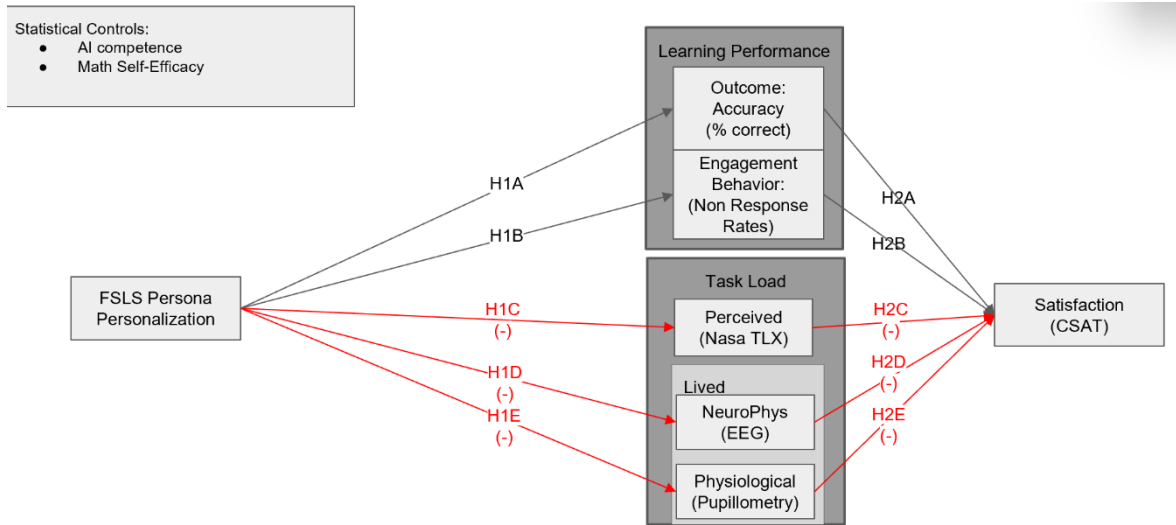
From a Cognitive Load Theory perspective, behavioural engagement is a visible manifestation of germane processing, the productive allocation of working memory resources to schema construction and automation (Sweller, 2019). When learners actively attempt difficult problems rather than disengaging (low nonresponse rates), or when they manipulate symbolic notation during tutoring (high formula use), they signal investment of cognitive effort. These behaviours contrast with passive reception or avoidance strategies that conserve mental resources but yield shallow learning. Sustained engagement is a necessary precondition for schema formation, particularly in domains such as mathematics, where procedural fluency requires extensive deliberate practice (Ericsson et al., 1993). By examining both behavioural engagement (process) and accuracy (product), we gain insight into whether personalization altered how learners interacted with mathematics content, independent of what they ultimately learned.

Overall, the reviewed literature indicates that persona-level adaptation in AI-mediated tutoring can affect not only learning outcomes but also the cognitive and behavioural processes that support learning. This provides the basis for the hypotheses that follow.

3. Hypothesis Development

We propose using learners' Felder-Silverman profiles to create a persona-level adaptation which enhances learning effectiveness and user experience while reducing task load. This overarching hypothesis is examined across three domains: learning outcomes, task load, and the latent construct of satisfaction, as shown in Figure 1.

Figure 1. *Research Model and Conceptual Framework of Effects of Personalized AI-Mediated Learning on Task Load and User Experience*



3.1 Learning Outcomes

Learning performance reflects the extent to which instructional personalization translates into measurable behavioural gains. Within AI-mediated tutoring, aligning instructional delivery with learners' processing preferences should enhance comprehension and encourage active engagement with mathematical content. According to Cognitive Load Theory, aligning instruction with learners' processing preferences reduces extraneous load and supports efficient schema construction (Sweller, 2019). Consequently, FSLS-based personalization is expected to improve both learner response accuracy and efficiency, as reflected in higher learning accuracy and lower nonresponse rates. Therefore, we posit the following hypotheses:

H1A: FSLS-based personalization will improve learning accuracy (percentage correct) compared to the control condition.

H1B: FSLs-based personalization will improve learning efficiency, operationalized as reduced non-response rates, compared to the control condition.

3.2 Task Load

According to CLT, extraneous cognitive load arises when instructional design imposes unnecessary processing demands (Sweller et al., 2011). Misalignment between the tutor's explanatory style and the learner's processing preferences constitutes a form of extraneous load: learners must mentally translate or reorganize information to align with their preferred format. FSLs-based personalization should eliminate this translation step, thereby reducing both subjective and objective indices of task load.

Neurophysiological evidence supports this prediction. Frontal theta power increases with working memory demands and sustained attention, while parietal alpha power decreases with cognitive fatigue and disengagement (Klimesch, 1999, 2012). Personalization that reduces extraneous load should be reflected in lower theta-alpha ratios during tutoring interactions. Similarly, reduced pupil dilation reflects cognitive effort under personalized conditions, indicating decreased momentary cognitive effort. Taken together, these considerations lead to the following hypotheses:

H1C: FSLs-based personalization will reduce perceived task load (NASA-TLX mental demand and frustration) compared to the control condition.

H1D: FSLs-based personalization will reduce neurophysiological task load (EEG theta-alpha ratios) compared to the control condition.

H1E: FSLs-based personalization will reduce physiological task load (pupil dilation) compared to the control condition.

3.3 User Satisfaction

User satisfaction with educational technology reflects the learner's holistic experience, influenced by objective performance and perceived ease of learning (Bolliger & Martindale, 2004; Martin & Bolliger, 2018). Learners who achieve higher accuracy and lower non-response rates, indicating successful learning, should report higher satisfaction. Conversely, learners who experience high cognitive load, whether measured subjectively or physiologically, should report lower satisfaction, as high load signals an effortful and potentially frustrating learning experience.

Accordingly, we propose the following hypotheses:

H2A: Learning accuracy (percentage correct) will positively predict satisfaction.

H2B: Non-response rates will negatively predict satisfaction.

H2C: Perceived task load (NASA-TLX) will negatively predict satisfaction.

H2D: Neurophysiological task load (EEG) will negatively predict satisfaction.

H2E: Physiological task load (pupil dilation) will negatively predict satisfaction.

4. Methods

4.1 Participants

Forty-five participants were recruited from a university research panel. Data from 40 participants were retained for analysis after excluding incomplete or compromised recordings. Participants ranged in age from 18 to 45 years ($M = 27$), with 29/40 identifying as female. Thirty-one participants completed a bachelor's degree or post-graduate degree, as shown in Table 2 when both segments are summed. All participants provided written informed consent prior to participation and were reminded that they could withdraw at any time without penalty. Participants received CAD \$60 for their participation, which lasted 2 hours. Ethics approval for the study was obtained from the institutional research ethics board (Certificate #2023-5071). Eligibility criteria ensured compatibility with the electroencephalography (EEG) system, including an appropriate head cap fit and scalp accessibility, whereas exclusion criteria accounted for factors such as excessive hair density or EEG-related health contraindications.

Table 2. *Participant Demographics: Age, Sex, and Education Level*

Variable	Category	<i>N</i> (%)	<i>M</i> (<i>SD</i>)	Median	Range
Age	-	<i>N</i> =40	27 (6.08)	25.5	18 to 45
Sex		<i>N</i> =40			
	Male	11 (27.5)	-	-	-
	Female	29 (72.5)	-	-	-
Education		<i>N</i> =39			
	High School	5 (12.8)	-	-	-
	Some College	3 (7.7)	-	-	-
	Bachelor's degree	17 (43.6)	-	-	-
	Post-Graduate Degree	14 (35.9)	-	-	-

Note. Categorical and continuous demographic data types are included in Table 2. One participant did not disclose their level of Education, which is why $N=39$.

4.2 Study Design

The study employed a mixed-subjects $2 \times 2 \times 3$ factorial design, illustrated in Table 3, followed by post-test assessments after each task. At the highest between-subject level of the factorial design, participants were randomly assigned to one of two learning-style groups derived from the Felder-Silverman Learning Style Inventory (FSLI), representing either the perception dimension (intuitive/sensing; Group A) or the understanding dimension (global/sequential; Group B), described in Table 4. This grouping ensured that personalization effects could be attributed to the experimental manipulation rather than to preexisting learning-style differences. At the second level, the primary manipulation concerned task-level personalization (treatment versus control), with each participant exposed to both conditions in a within-subjects design. Finally, at the third level, three blocks representing different mathematical topics were presented in counterbalanced order to mitigate sequence effects.

Each participant completed three blocks, each consisting of a tutoring interaction followed by a test on a different math topic. After each seven-minute tutoring session, participants completed a five-minute comprehension quiz.

Across the 40 participants, we collected 120 blocks of data (61 treatment; 59 control), using a within-subjects design to maximize statistical power. According to Cohen (1992), a minimum of 34 participants is required to achieve 80% power to detect a large effect ($\alpha = .05$) (Cohen, 1992). Our sample exceeded this threshold.

Table 3. *2X2X3 Factorial Design with Between- and Within-Subjects Comparisons*

Perceptual (Group A): Sensing/Intuitive (n=60)				
Condition	Math Subject			
		Venn Diagram (n=10)	Parabola (n=10)	Logarithms (n=10)
	Treatment (n=30)	A-T1	A-T2	A-T3
	Control (n=30)	A-C1	A-C2	A-C3
Organization (Group B): Global/Sequential (n=60)				
Condition	Math Subject			
		Venn Diagram (n=10)	Parabola (n=10)	Logarithms (n=10)
	Treatment (n=31)	B-T1	B-T2	B-T3
	Control (n=29)	B-C1	B-C2	B-C3

Note. Participants are organized into subgroups based on FSLs-Group (A or B), then each Task is assigned a condition (T or C), and a math topic (Venn, Parabola or Logs). These variables make up the three digit code in each block.

4.3 Materials

AI Tutor Platform

The AI tutor was implemented by integrating a custom front-end interface, an API gateway, and a large language model (LLM) programmed in Python version 3.11 (Python Software & Van Rossum, 2022). The custom platform was necessary because it permitted precise manipulation of the independent variable that defined the tutor’s persona. It also enabled complete logging of all user interactions and ensured consistent capture of query content, response content, timestamps, and metadata required for prompt-level analyses and information retrieval. The study used

Mistral’s Large 2 model (123B parameters), a refined version of the large series designed to minimize hallucinations, follow detailed instructions, and manage long interactions effectively (Mistral, 2024). It was chosen for its strong performance on mathematical reasoning benchmarks, which align with the study’s educational context. The interface provided a text-based chat environment allowing participants to submit queries and receive generated responses, with an adjustable font size for accessibility. Participants entered text using a keyboard and submitted queries by clicking a button with the mouse or pressing Enter. Responses could be re-sent if an error occurred, and researchers adjusted font size or styling to accommodate accessibility requests (e.g., one participant used a larger 14-point font) prior to the commencement of the task. See figure2 for an image of the chatbot interface.

System Prompts

System-level prompts defined the tutor’s persona and ensured consistency across conditions. A shared core-prompt established baseline parameters, including the chatbot’s role, interaction rules, formatting of mathematical expressions, and session scope. Experimental customization was achieved through additional instructions: (1) learning-style adaptation (see Table 4), which tailored responses to FSLS group assignment (e.g., step-by-step explanations for sequential learners versus high-level summaries for global learners), and (2) math subject and scope, which specified task content. In the control condition, the tutor operated with only the core prompt and topic-specific instructions, without learning-style adaptation. Access to the full system-level prompts is provided in Appendix C.

System Prompt Development

Core prompts were developed through literature review and iterative pre-testing. Research indicates that role-based personas (e.g., “You are a tutor”) have a minor effect on response accuracy, where extraneous details such as persona gender should be removed unless they provide specificity and match the task domain and type (e.g., “You are a **math** tutor **with knowledge in ...**”) (Zheng et al., 2024). During Pretests, we confirmed the above finding and the need to include explicit instructions for the AI to display equations, ensuring consistent notation across mathematical topics.

Learning-style adaptations were constructed from the FSLS dimensions (intuitive/sensing and sequential/global). Each dimension was mapped to the corresponding pedagogical strategies described by Felder and Silverman and then translated into system-prompt instructions, as illustrated in Table 4 (Felder, 1988; Haug et al., 2023). Pre-tests revealed that certain core instructions in the first segment of the prompt suppressed the intended downstream effects of the prompt's learning-style personalization; minor adjustments were therefore implemented to balance general consistency with style-specific adaptations.

Table 4. *Learning Styles Applied in the Experiment and Prompt Segments Used to Personalize the AI Math Tutor*

Perception Dimension (Group A): What kinds of information learners prefer		
learning Types	Description	Prompt used in AI design
Sensing	Prefer concrete, practical information and real-world examples. They like facts, details, and established methods, and they are often good with hands-on application.	"Start by explaining what this math concept is used for in real life, with a concrete example... Use clear, practical language. Avoid abstractions or metaphors unless they are tied to something physical or observable..."

Intuitive	Prefer abstract concepts, theories, and underlying meanings. They enjoy innovation, connections across ideas, and discovering relationships rather than following set procedures.	"Begin with an interesting idea, theory, or big question behind the topic... Emphasize connections between ideas... Use generalizations, theorems, and patterns rather than focusing solely on isolated cases. Offer multiple ways of approaching a concept to encourage creative thinking... Encourage the learner to make predictions or before giving the answers (e.g., "What do you expect to happen if we double this value?")."
Understanding/Organization Dimension (Group B): How learners build knowledge structures		
Sequential	Prefer linear, step-by-step learning. They progress in logical order, building understanding gradually.	"State the learning objective clearly and specifically. Present information in small, logical steps... Follow a linear progression, avoiding big jumps in reasoning... Provide clear transitions between stages of learning. Emphasize procedures, and step-by-step methods. Don't give all the instructions at the same time."
Global	Prefer holistic learning. They often absorb information in large chunks, sometimes appearing to learn suddenly after struggling. They look for the "big picture" first before details make sense.	"The user doesn't need to understand the concept in detail, it's just so they are situated mentally... write the overall goal of the lesson, the general concept. Once the overall goal has been presented, go a little more in details. You can name all the concepts listed in your prompt at the same time and give a brief description, so the person understands the direction they're going... progressively go into more detail."
Processing Dimension (Not included): How learners engage with information		
Active	Prefer learning by doing through engagement, discussion, group work, and hands-on practice. They tend to retain and understand material best when they try it out or explain it to others.	NA
Reflective	Prefer thinking through material quietly before acting. They benefit from reviewing, summarizing, and reflecting individually rather than jumping straight into action.	NA
Input Dimension (Not Included): How information is best received		
Visual	Learn best from pictures, diagrams, flowcharts, videos, and other graphical representations of information.	NA
Verbal	Learn best from words. Spoken explanations, written text, discussions, and listening to others.	NA

Finally, introductory system messages were standardized across conditions, with slight topic-dependent variations. Scope limitations ensured that each tutor covered only the designated

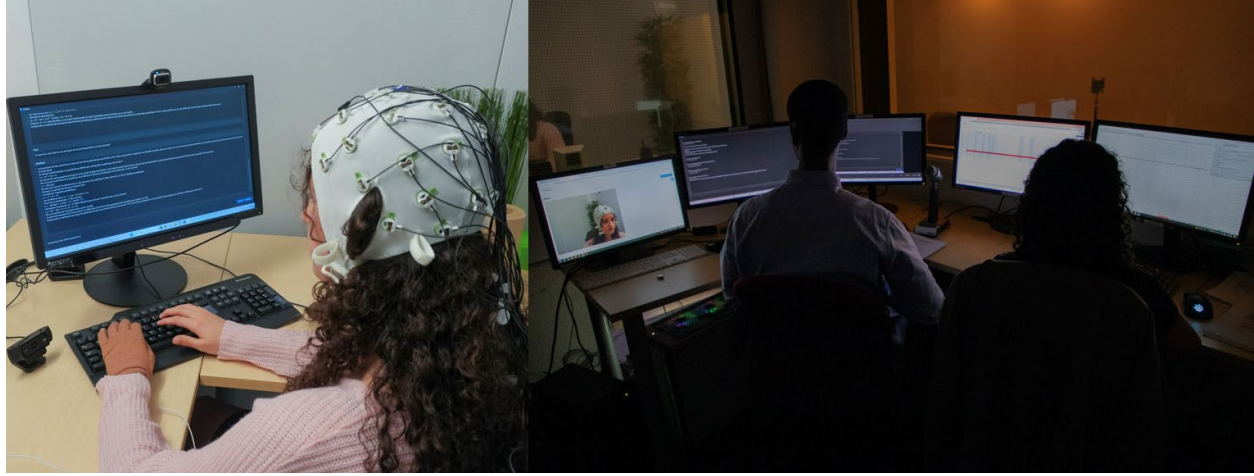
mathematical content, preventing the student from deviating from the content covered in the post-task assessment.

4.4 Procedure

Data collection occurred in June 2025. Each participant completed a single 2.5-hour session in a Faraday cage designed to minimize electromagnetic and acoustic interference. The room was equipped with a height-adjustable chair and a Lenovo ThinkVision monitor (1920×1080 at 59.93 Hz), as shown in Figure 2 (left panel).

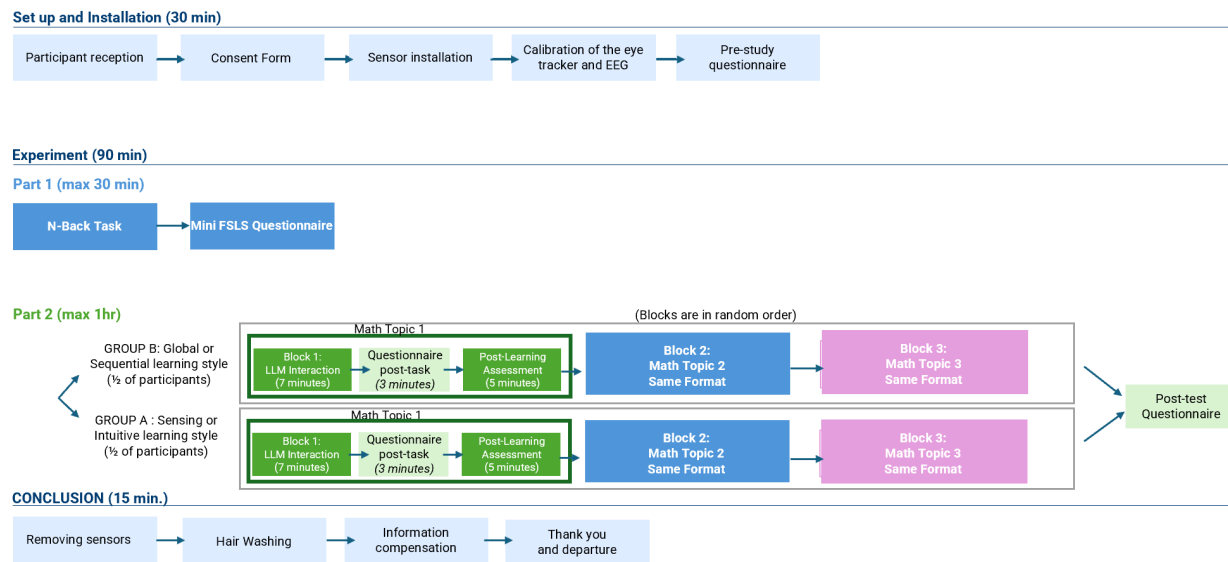
The study used two adjoining rooms separated by a one-way mirror. Participants were seated alone, while the research team observed from the control room, which contained three computers and five monitors: one for webcam recordings of facial expressions, one for presenting experimental stimuli, and one for EEG recording (see right image of Figure 2). Researchers communicated with participants via intercom. Lighting and temperature were kept constant across sessions, and both electromagnetic and acoustic interference were minimized.

Figure 2. *Laboratory Design Demonstration: Participant View, Lab Assistant Demonstration, and Observation Room Setup*



Experimenters adhered strictly to standardized protocols, delivering consistent oral and written instructions to participants. Continuous monitoring ensured correct EEG operation, with manual timestamp verification and alignment conducted in real time.

Figure 3. *Experimental Design Including Steps and Duration of Each Phase*



The subsequent subsection provides a detailed description of the procedures implemented during the experiment. As shown in Figure 3, the study followed a structured sequence of steps, and the subsequent text elaborates on these steps to document the full flow of the experimental protocol.

Before the experimental tasks, participants read and signed an informed consent form and were reminded of their right to withdraw at any point without penalty. Following consent, participants completed a short oral questionnaire to gather demographic information.

Following the demographic questionnaire, the EEG cap (actiCAP Slim 32, BrainProducts GmbH, Germany) was fitted to each participant's head. Head circumference was measured to select the correct cap size. A conductive gel was applied to each electrode site, and impedances were verified to be below the manufacturer's and laboratory thresholds. Reference and ground electrodes were positioned according to the 10-20 system (Jasper, 1958). EEG signals were continuously recorded at 500 Hz using the BrainVision Recorder software.

The eye-tracking system (Tobii Fusion Pro, 60 Hz) was then calibrated using a standard nine-point calibration procedure in Tobii Pro Lab. Participants were seated approximately $65 \text{ cm} \pm 5 \text{ cm}$ from the monitor to ensure optimal gaze detection and comfort. A one-minute baseline of resting pupil diameter was recorded prior to any task onset, providing a reference for later normalization of task-evoked pupil responses.

Participants first completed a pre-test questionnaire, via Qualtrics, to collect data on the control variables: educational background, AI literacy (Wang et al., 2023), and math self-efficacy (May, 2009). Then participants completed an N-back working memory task (PsychoPy) to obtain a

baseline estimate of individual working memory capacity and task-related EEG dynamics.

Participants were instructed to press a designated key when the current letter matched the one presented N positions earlier in the sequence. Four levels (0-, 1-, 2-, 3-back) were administered in ascending order of difficulty. Reaction times and accuracy were recorded, and concurrent EEG provided a participant-specific calibration of cognitive load under known demands.

Participants then completed an abridged 11-item Felder-Silverman Learning Style (FSLs) questionnaire, later used to personalize the AI-tutor interactions.

Following calibration, participants entered the learning phase, comprising three tutoring blocks. Each block corresponded to a distinct mathematics domain: (1) probability: Venn diagrams, (2) quadratic functions: parabolas, and (3) logarithmic functions. Block order was counterbalanced across participants to mitigate topic-order effects.

Each block comprised two segments for all participants, regardless of the assigned condition (control or treatment). The first segment is the Tutoring interaction (7 minutes), where participants interacted via text chat with an AI tutor. They were instructed to ask any clarifying questions necessary to understand the topic. The interface displayed the conversation window. When the segments are assigned as a treatment condition, the tutor's system prompt includes learning-style adaptations based on the participant's Felder-Silverman Learning Style (FSLs) profile (as seen in Table 4), derived from their pre-experiment inventory. In control blocks, the tutor included the same topic content but without any personalization parameters, maintaining only the baseline instructional script. EEG and eye-tracking data were continuously recorded

during the entire 7-minute tutoring interval. No external aids (e.g., calculators, notes) were permitted.

Immediately following each tutoring session, participants completed a short post-interaction questionnaire that assessed perceived task load, satisfaction, and a manipulation-check question on perceived personalization (see Table 5).

Following the post-interaction questionnaire, participants completed a multiple-choice quiz assessing conceptual comprehension with a maximum time of 5 minutes. Each quiz included four response options per item, one of which was “I do not know,” to distinguish uncertainty from incorrect knowledge. Quizzes were presented on-screen, and responses were logged automatically. At the end of the session, all electrodes were removed, and residual gel was cleaned.

4.5 Data Acquisition

Electrophysiological and ocular data were continuously recorded throughout the experiment. Electroencephalographic (EEG) activity was measured using a 32-channel actiCAP slim system with active electrodes and a LiveAmp32 wireless amplifier (BrainProducts GmbH, Munich, Germany). Signals were digitized at 500 Hz and recorded using BrainVision Recorder software. Electrodes were positioned according to the international 10-20 system, with the ground electrode placed at AFz and the reference at FCz. Impedances were maintained below 10 k Ω for all channels prior to data collection. For subsequent analyses, eight electrodes of interest were retained: frontal (Fp1, Fp2, F3, F4) and parietal (CP1, CP5, P3, Pz), corresponding to regions

implicated in attention, working memory, and cognitive control (Cavanagh & Frank, 2014; Klimesch, 1999; Sollazzo et al., 2025; Wascher et al., 2014).

Eye movements and pupil responses were recorded using a Tobii Fusion Pro eye tracker (Tobii Pro, Stockholm, Sweden) operating at 60 Hz. A standard nine-point calibration was completed using Tobii Pro Lab software before each session. Participants were seated approximately 65 ± 5 cm from the display monitor, within the optimal range for accurate gaze detection. A one-minute baseline recording of resting pupil diameter was obtained prior to task onset to serve as a reference for later normalization.

Behavioural data, including quiz responses, response times, and N-back accuracy, were logged through PsychoPy and the AI tutor interface. Questionnaire data were collected through Qualtrics (Provo, UT, USA). To ensure temporal alignment, all systems were synchronized using manual timestamp verification and logged event markers. Experimenters continuously monitored signal integrity and synchronization from the control room and manually applied timestamps to mark transition events for subsequent processing.

4.6 Preprocessing and Feature Extraction

EEG data were preprocessed in two stages using different software tools. First, using BrainVision Analyzer (BrainProducts GmbH) to insert event markers, following standard laboratory procedures. The second preprocessing stage was conducted using MNE-Python, in which signals were subjected to band-pass filtering (0.5-50 Hz) to remove slow drift and high-frequency noise, and to a notch filter (58-62 Hz) to eliminate electrical interference (Gramfort et

al., 2013). Artifacts arising from blinks or eye movements were corrected through independent component analysis (ICA), followed by visual inspection to reject residual contaminated segments. The cleaned data were segmented into six-second epochs corresponding to the tutoring and post-test periods, as defined by event markers.

For each epoch, the power spectral density (PSD) was estimated using Welch's method with 50% overlap and Hamming windows. Mean spectral power was extracted for the theta (4-7 Hz) and alpha (8-13 Hz) bands. Task-load indices were computed as ratios of frontal theta to parietal alpha power. Sixteen pairwise ratios were generated by combining four frontal (Fp1, Fp2, F3, F4) with four parietal (CP1, CP5, P3, Pz) electrodes, following the approach validated by Sollazzo (2025). Ratios were averaged across epochs within each experimental block, producing a single frontal-parietal task-load index per condition and participant.

Pupillometric data were processed in Tobii Pro Lab. Blink-related gaps were automatically corrected using moving filters as part of Tobii Pro Lab software features. Baseline correction was applied by subtracting the mean pupil diameter recorded during the baseline calibration from each subsequent value. Mean pupil dilation was then computed for each six-second epoch and temporally aligned with the corresponding EEG epochs.

4.7 Measures

This study assessed cognitive, behavioural, linguistic, and subjective outcomes using neurophysiological recordings, self-report instruments, performance-based indicators, and interaction-level transcript analyses. The constructs included learning performance, behavioural

engagement, linguistic elaboration during tutoring, perceived task load, physiological task load, satisfaction, perceived personalization, and individual difference controls. Table 5 provides a structured overview of these measures, including item formats, response scales, and reliability coefficients.

The behavioural performance measures comprised accuracy on the post-learning assessments, completion rate, nonresponse rate, the proportion of “I do not know” selections, and the correct-if-answered rate. These indicators provided complementary information about performance tendencies and response strategies during the math tasks.

Subjective experience was measured through multiple self-report scales. Satisfaction was assessed with a single-item Customer Satisfaction Score (CSAT) commonly used in service evaluation contexts (Farris et al., 2010). Perceived personalization was assessed with a manipulation check that evaluated whether the tutor appeared to tailor its explanations to the user’s needs. Perceived task load was captured using the mental demand and frustration subscales from the NASA Task Load Index, established indicators of subjective workload (Hart & Staveland, 1988).

As previously noted, task load refers to the learner's cognitive load during the immediate learning activity, excluding factors such as emotional arousal and the learner's environmental context (Sollazzo et al., 2025). This narrow operationalization aligns with the focus of CLT and Dual Process Theory (DPTs) on the immediate cognitive costs of information processing (Evans & Stanovich, 2013; Sweller et al., 2011). Task load was assessed using three complementary methods that combined physiological and self-report measures.

Neurophysiological task load was measured using electroencephalography (EEG) and pupillometry. The EEG index was based on ratios of frontal theta to parietal alpha power, computed across sixteen electrode pairings (Rozado & Dunser, 2015; Sollazzo et al., 2025). Pupillometry measured task-evoked pupil dilation relative to baseline as an indicator of arousal and cognitive effort (Krejtz et al., 2018; Rozado & Dunser, 2015).

Interaction-level measures were derived from transcripts of the AI tutor exchanges. These measures captured behavioural engagement during each seven-minute tutoring block. Specifically, the total number of prompts sent by the user and the number of responses generated by the AI tutor were computed for each block. Additionally, the proportions of user prompts and tutor responses containing mathematical equations were analyzed.

Two individual difference constructs served as statistical controls. Artificial Intelligence Competence was measured using a validated five-item scale assessing users' perceived ability to understand and use AI applications (Wang et al., 2023). Math Self-Efficacy was measured with a six-item questionnaire capturing confidence and anxiety related to mathematical tasks (May, 2009)

Together, these measures provided a multimodal assessment of learning processes during AI-mediated tutoring. Full scale administration details such as wording, response formats, and scale characteristics appear in Table 5.

Table 5. *Description of Constructs and Measurement Tools Along with Cronbach's Alpha*

Construct Name	Measure Type	Measurement Tool	Description	Source reference	Cronbach's Alpha
AI-Competence (control)	Self-reported/psychometric	Questionnaire (5 items)	1. I can distinguish between smart devices and non-smart devices. 2. I can skillfully use AI applications or products to help me with my daily work. 3. I can use AI applications or products to improve my work efficiency. 4. I can choose a proper solution from various solutions provided by a smart agent. 5. I can choose the most appropriate AI application or product from a variety for a particular task. Scale response options: 1 = strongly disagree 7 = strongly agree	Wang, B., Rau, P. L. P., & Yuan, T. (2022). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. <i>Behaviour & Information Technology</i> , 42(9), 1324–1337.	0.846
Math Self-Efficacy (control)	Self-reported/psychometric	Questionnaire (6 items)	1. I believe I am the kind of person who is good at mathematics. 2. I feel confident when taking a math test. 3. I'm certain I can understand the most difficult material presented in math texts. 4. I worry that I will not be able to do well on mathematics tests. 5. I get nervous when I have to use mathematics outside of school. 6. I get nervous when taking a mathematics test. Scale response options: 1 = strongly disagree 7 = strongly agree	May, D. K. (2009). <i>Mathematics Self-Efficacy and Anxiety Questionnaire</i> [Doctoral dissertation, The University of Georgia]. Athens, Georgia.	0.917
Satisfaction	Self-reported/psychometric	Questionnaire (1 item)	CSAT Are you satisfied with the AI Math tutor? Scale response options: 1 = Very Dissatisfied 7 = Very Satisfied	Farris, P. W., Bendle, N., Pfeifer, P., & Reibstein, D. (2010). <i>Marketing Metrics: The Definitive Guide to Measuring Marketing</i>	

				Performance. Pearson Education.	
Task Load	Neurophysiological	EEG & Pupillometry	Physiologically there are two means of measuring task load. 1. EEG: Frontal Theta / Parietal alpha ratios from 16 pairings. 2. Pupillometry: Change in pupil diameter size	(Sollazzo, 2025) (Krejtz et al., 2018) (Rozado & Dunser, 2015)	
Perceived Task Load	Self-reported/psychometric	Questionnaire (2 items)	1. How mentally demanding was the task? - Mental Demand 2. How insecure, discouraged, irritated, stressed, and annoyed were you? - Frustration Scale response options: percentage slider 0=low, 100=high From NASA-TLX	Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. <i>Advances in Psychology</i> , 52, 139-183.	0.785
Manipulation Check	Self-reported/psychometric	Questionnaire (1 item)	To assess whether users perceived a difference across tasks as they switched between control and treatment. Rate your level of agreement with the following statement: The chatbot tailored its explanations to my individual learning needs.” Scale response options: 1 = strongly disagree 7 = strongly agree		
Accuracy	Measured	Participant Assessment (12-16 questions)	The % correct answers on post-tests.		
Completion Rate	Measured	Participant Assessment (12-16 questions)	rate of questions answered		
Nonresponse rate	Measured	Participant Assessment (12-16 questions)	% of questions that were either left blank or marked as "I don't know"		
IDK rate	Measured	Participant Assessment (12-16 questions)	% of responses marked as "I don't know"		

Proportion of prompt that includes an equation	Measured	AI-Interaction Transcripts	The proportion of prompts sent by the user that include an equation within a 7-minute interaction with the LLM.		
Proportion of LLM response that includes equation	Measured	AI-Interaction Transcripts	The proportion of AI tutor responses that include an equation within a 7-minute interaction with the LLM.		
Number of user prompts sent per interaction	Measured	AI-Interaction Transcripts	Total number of messages that users sent within a 7-minute interaction with the LLM.		

4.8 Statistical Analysis

Analyses were conducted in SAS Online (Institute, 2024). Prior to inferential testing, psychometric scale items were reduced by examining factor loadings and internal reliability (Cronbach's $\alpha > .70$). Initial assessment of the dataset confirmed that there were sufficient data points to justify the use of parametric tests within a frequentist framework.

Following guidelines from the UX Data Analysis Toolkit (Fredette, 2025), the primary analyses employed linear mixed-effects models with random intercepts and repeated measures to account for the within-subjects structure. Models tested the effects of personalization condition, task order, and learning-style group on learning and experiential outcomes. Secondary analyses included repeated-measures ANOVAs for tests with more than 2 categorical variables, for example, the task-level analysis (Tasks 1-3). P-values from single-tailed hypothesis tests were divided by 2; this procedure includes all models listed in Table 6, which reports the primary hypothesis tested. To control for inflation of Type I error due to multiple comparisons, we adjusted p-values post hoc using Holm's step-down procedure for a pairwise analysis, and we

corrected confidence intervals using the Bonferroni method. Control variables were included if the interactions were significant (see Table 9).

Likert-scale data were treated according to the design: five-point scales were treated as ordinal and seven-point scales as continuous, based on distributional properties. Assumptions of normality and homogeneity of variance were tested, and models were adjusted when necessary. Statistical significance was set at $\alpha = .05$, with exploratory analyses reported at $\alpha = .10$ to balance statistical rigour with sensitivity to emerging effects. This approach reduces the risk of Type II errors and aligns with other experimental neuroscience research that treats marginally significant findings ($0.05 < p < 0.10$) as indicative of meaningful trends (Healy et al., 2015).

5. Results

Four sets of analyses were conducted to examine the effects of personalization on learning and task-related outcomes. The main comparisons focused on differences between the personalized (treatment) and non-personalized (control) conditions (H1) and on associations between the measured variables and participants' reported satisfaction (H2), as presented in Table 6.

Table 6. *Hypothesis Test Results*

Hypothesis	From	To	β	t	p	Status
H1A	Personalization	Accuracy (% correct)	-0.005	0.198	0.422	Not Supported
H1B	Personalization	Efficiency (Non-response Rate)	-0.052	1.516	0.095*	Supported at $\alpha = 0.1$
H1C	Personalization	Task Load (TLX)	2.917	-0.951	0.345	Not Supported
H1D	Personalization	Task Load (EEG)	-0.502	5.42	< 0.001 ***	Supported

H1E	Personalization	Task Load (Pupil Size)	-0.055	2.81	0.006 ***	Supported
H2A	Accuracy (% Correct)	Satisfaction	1.919	2.93	0.002 ***	Supported
H2B	Non-response Rate	Satisfaction	-1.491	-2.78	0.003 ***	Supported
H2C	Task Load (TLX)	Satisfaction	-0.03	-5.91	< 0.001 ****	Supported
H2D	Task Load (EEG)	Satisfaction	-0.073	-1.02	0.157	Not Supported
H2E	Task Load (Pupil Size)	Satisfaction	0.449	0.86	0.196	Not Supported

Note. Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is *** $p < .001$ is ****

To ensure the robustness of the effects from Table 6, two additional exploratory comparisons were included: differences across the three successive tasks (Table D2 in Appendix D) to observe an ordering effect and control for task-specific variability, and differences between learning style groups A and B to account for differences specific to the types of learning style personalization's (Table D3 in Appendix D), both set of supplementary results are reported here and further elaborate tables were placed in the Appendix D. Together, these analyses provide a comprehensive assessment of how personalization influenced behavioural, subjective, and physiological measures of learning outcomes.

5.1 Effects of Personalization

To address Hypotheses 1A to 1E, we examined whether personalization produced measurable differences across behavioural, perceptual, and physiological outcomes. Table 7 presents descriptive and inferential statistics comparing the treatment and control conditions for each dependent variable. These analyses provide the primary test of the effects of persona-level

personalization on task performance and task load. The results for each outcome are summarized below.

Table 7. *Linear Mixed-Effects Results Comparing FSLSM Personalization and Control Conditions*

Variables	Control (<i>M</i>)	Control (<i>SD</i>)	Treatment (<i>M</i>)	Treatment (<i>SD</i>)	<i>df</i>	<i>F</i>	<i>p</i>
Accuracy	0.38	0.19	0.37	0.20	84.53	0.04	0.422
Nonresponse rate	0.31	0.26	0.28	0.23	78.74	1.74	0.095*
Completion rate	0.93	0.12	0.95	0.10	89.67	1.22	0.136
Rates of "I don't know"	0.24	0.25	0.22	0.23	77.14	0.26	0.306
Perceived task load	48.80	24.61	52.43	24.34	82.89	0.90	0.172
EEG task load (AI interaction)	2.69	2.94	1.76	1.72	71.21	29.35	<.000****
EEG task load (math test)	2.65	2.96	1.72	1.79	71.49	10.09	.001****
Pupil task load (AI interaction)	0.82	0.36	0.79	0.31	73.52	7.91	.003***
Pupil task load (math test)	-0.02	0.20	-0.0001	0.21	74.82	0.05	0.408
Number of prompts sent per interaction	6.24	2.73	5.41	2.61	83.98	2.02	0.080*
Proportion of prompt with equation	0.01	0.04	0.03	0.08	92.45	2.48	0.059*
Proportion of LLM responses with equation	0.59	0.28	0.65	0.29	91.01	0.71	0.200
Manipulation Check	4.37	1.52	4.41	1.69	83.55	0.03	0.862
Satisfaction	4.36	1.35	4.43	1.44	85.11	0.23	0.315

Note. Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is *** $p < .001$ is ****. *DF* = Degrees of Freedom

5.2 Learning Gain

Comparisons of learning outcomes between the personalized (treatment) and control conditions using a linear mixed-effects model with random intercepts revealed no significant differences in accuracy ($F(1, 84.53) = 0.04, p = .422$). Nonresponse rates, a composite of blank and “I do not know” answers, were lower for the personalized condition ($\beta = -0.052, t(78.74) = 1.32, F(1,$

78.74) = 1.74, $p = .095$), although marginally significant when $\alpha = .10$. Completion rate and IDK rate did not differ significantly across conditions.

5.3 Satisfaction

We used a linear mixed-effects model with random intercepts for the series of Satisfaction tests (H2-series). Satisfaction ratings were strongly predicted by subjective task load. NASA-TLX scores negatively predicted satisfaction ($\beta = -0.030$, $t(114.1) = -5.91$, $F(1, 114.1) = 34.96$, $p < .001$), confirming that participants who reported greater workload experienced lower satisfaction. Higher accuracy significantly predicted higher satisfaction ($\beta = 1.92$, $t = 2.93$, $p = .004$), whereas greater nonresponse rates significantly predicted lower satisfaction ($\beta = -1.491$, $t = -2.78$, $p = .003$).

5.4 Task Load

5.4.1 Perceived

No significant main effect of personalization on perceived task load was observed ($\beta = 0.35$, $p = .546$). Mean workload ratings were comparable between personalized and control conditions.

5.4.2 Pupillometry

Personalization significantly reduced physiological task load during the tutoring session. Results shown in Figure 5 indicate smaller average pupil dilations compared with the control condition ($\beta = 0.055$, $t(73.52) = 2.81$, $F(1, 73.52) = 7.91$, $p = .003$). Task-wise analyses, using a linear

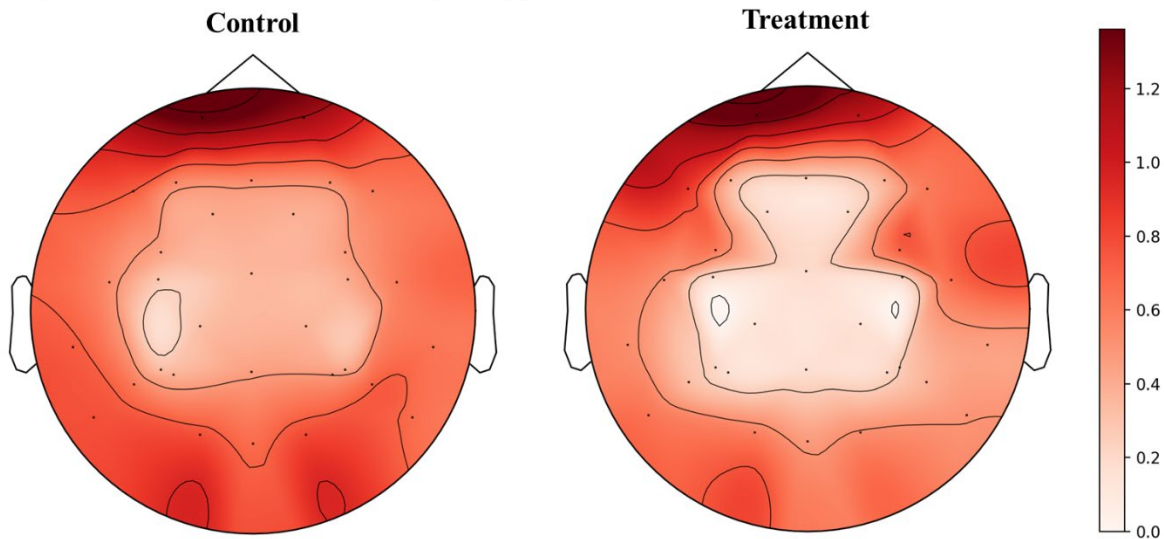
mixed-effects logistic regression with Holm's adjustment for p-values, revealed a significant main effect of task order on pupil size, $F(2, 72) = 8.32$, $T1p = .0002$. Pupil size decreased across blocks ($M^{T1} = 0.85$, $M^{T2} = 0.79$, $M^{T3} = 0.77$) during the interactions with significant reductions between Task 1 and Task 2 and between Task 1 and Task 3 (Figure 6, Panel B); however, there was no significant change in pupil size during the post-test.

In the multi-group comparison of control v. group A v. group B personalization, again using a linear mixed effects ANOVA with Holm's adjustment, the mean pupil dilation from baseline during the tutoring session was smallest in Group B ($M = 0.803$) and largest in the control condition ($M = 0.821$), with Group A in between ($M = 0.772$). Although the overall effect across groups approached significance, $F(2, 74.82) = 4.03$, with a global $p = .095$, the comparison that was closest to being significant was the holms adjusted pairwise comparison of Control and Group A ($F(2, 74.87) = 4.03$, $p = .072$) at $\alpha = .10$.

5.4.3 EEG

EEG indices confirmed the reduction in neural effort under personalization. During the learning phase, frontal-theta to parietal-alpha ratios were significantly lower in the personalized condition than in control ($\beta = -0.502$, $t(71.21) = 5.42$, $F(1, 71.21) = 29.35$, $p < .001$). The same effect persisted during the subsequent test phase ($\beta = -0.455$, $t(71.49) = 3.18$, $F(1, 71.49) = 10.09$, $p = .001$), demonstrating sustained cognitive-efficiency gains following adaptive interaction.

Figure 4. *Topographic Distribution of Theta Band Power During Tutoring Tasks*



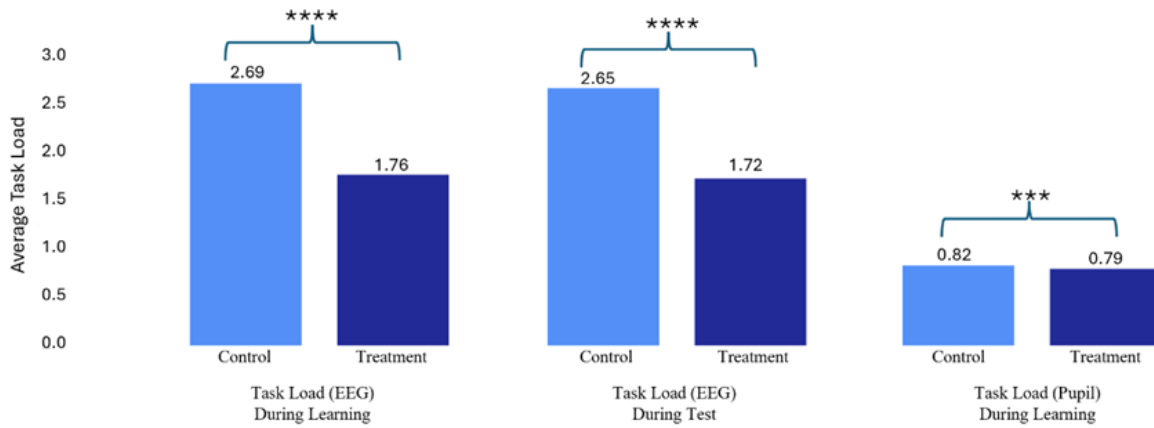
Note. Time frequency topographic maps depicting average theta band power (4 to 7 Hz) during the control and treatment conditions. Darker colours indicate higher theta power. Maps are presented using the standard 10-20 system front orientation with frontal sensors at the top.

Figure 4 presents the spatial distribution of theta band power across the scalp during the tutoring tasks. The treatment condition shows lower theta power in the frontal regions (Fp1 and Fp2).

Reduced frontal theta is associated with lower overall task load.

In the multi-group comparison (Group A vs B vs control) using a linear mixed-effects ANOVA with Holm's correction (Table 3D in Appendix D), the EEG-derived task-load ratio differed significantly across personas, $F(2, 71.17) = 15.59, p = .006$. The control condition exhibited the highest average task-load value ($M = 2.69$), followed by Group A ($M = 1.97$) and Group B ($M = 1.53$). Pairwise contrasts showed that both Group A and Group B differed significantly from the control group, with a larger reduction in task load for Group B.

Figure 5. Average Task Load Across Conditions Measured by EEG and Pupillometry



Note. Bars represent the mean task load for the Control and Treatment tutors during the learning and test phases. Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is *** $p < .001$ is ****

5.5 Control Variables

We tested whether the effects of the tutoring condition varied as a function of the control variables, including current level of education, self-reported math self-efficacy, and AI-competence. None of these interaction terms were significant, as shown in Table 8, indicating that the treatment effects were not dependent on these learner characteristics.

Table 8. Statistical Control Variable Screening Results from Linear Regression of Personalization Effects on EEG-Derived Task Load

Control Variable	Interaction with Variable	p	Retained in Analysis
Education Level	Personalization X EEG (TL)	0.615	No
Math Self-Efficacy	Personalization X EEG (TL)	0.074	No
AI-Competence	Personalization X EEG (TL)	0.070	No

Note. Results from individual multiple linear regression models to check for interaction effects. Control variables with P-values > 0.05 were not retained in the final models analyzed in the results.

5.6 Triangulation

Correlation analyses in Table 9 indicate limited convergence among subjective, pupillometric, and EEG indices. Perceived task load (NASA-TLX) scores were not correlated with either physiological metric. EEG-based task load during the tutoring interaction was weakly but significantly correlated with pupillometric task load ($r = -0.195, p = .046$), and EEG-based task load during the post-test showed a similarly weak correlation with pupillometric task load during tutoring ($r = -0.254, p = .009$).

Table 9. *Correlation Results Between Satisfaction and Dependent Variables*

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Accuracy	—	r	r	r	r	r	r	r	r	r	r	r
2. Nonresponse rate	-.81***	—	r	r	r	r	r	r	r	r	r	r
3. Completion rate	0.09	-.29**	—	r	r	r	r	r	r	r	r	r
4. Rates of “I don’t know”	-.79***	.90***	0.16	—	r	r	r	r	r	r	r	r
5. Perceived task load	-.27**	.19*	-.03	.19*	—	r	r	r	r	r	r	r
6. EEG task load (during AI interaction)	0.02	-.06	0.03	-.05	0.02	—	r	r	r	r	r	r
7. EEG task load (during math test)	0.03	-.08	0.02	-.08	0	.95***	—	r	r	r	r	r
8. Pupil task load	.19*	-.08	-.13	-.14	0.03	-.20*	-.25**	—	r	r	r	r
9. Proportion of prompt with equation	0.16	-.02	-.02	-.05	0	-.04	-.02	0.07	—	r	r	r
10. Proportion of LLM responses with equation	-.13	0.07	0.03	0.08	0.08	-.03	-.04	-.06	0.1	—	r	r
11. Manipulation Check	0.16	-.12	-.06	-.15	-.08	-.13	-.12	.20*	0.03	-.16	—	r
12. Satisfaction	.32***	-.30**	-.04	-.33***	-.47***	-.11	-.03	0.07	-.01	-.14	.54***	—

Note. Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is ***, $p < .001$ is ****

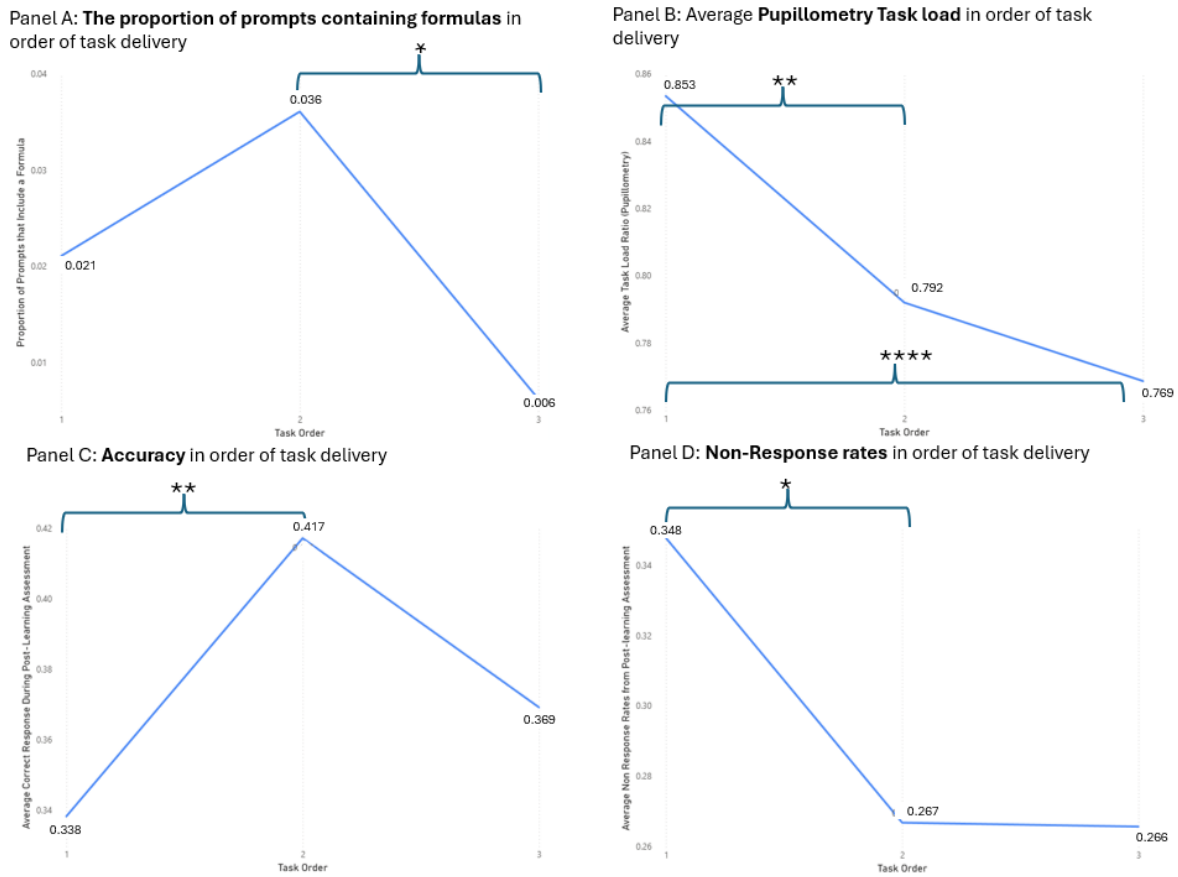
5.7 AI Tutor Interactions

Analyses of dialogue behaviour revealed effects of personalization and task progression.

Personalization increased the proportion of participant prompts containing mathematical formulas ($\beta = 0.018$, $F(1, 92.45) = 2.48$, $p = .057$, one-tailed) and decreased the number of prompts sent per interaction ($\beta = -0.523$, $F(1, 83.98) = 2.02$, $p = .040$, one-tailed). Task-order analyses showed a significant main effect of task on formula use ($F(2, 78) = 2.42$, $T2\ p = .031$): the proportion of formula-containing prompts increased from Task 1 ($M = 0.02$) to Task 2 ($M = 0.04$) and then declined in Task 3 ($M = 0.01$), to a significant degree, although marginal ($p = 0.0928$), as seen in Figure 6, Panel A.

Across several variables, there are significant changes from Task 1 to Task 2, and on one occasion from Task 2 to Task 3. Adjusted using Holms method, Nonresponse rate (Figure 6, Panel D) decreased from T1 to T2 ($M^{T1}=0.35 \rightarrow M^{T2}=0.27 \rightarrow M^{T3}=0.27$; $F(2, 74.71) = 2.89$, $p = .061$), and tutoring pupil dilation (Figure 6, Panel B) decreased ($M^{T1}=0.85 \rightarrow M^{T2}=0.79 \rightarrow M^{T3}=0.77$; $F(2, 72) = 8.32$, $p < .001$). These results show a large difference in performance and behavioural metrics between the first and second learning blocks, with subsequent stabilization in block 3.

Figure 6. Behavioural and Physiological Metrics Across Tasks in Chronological Order

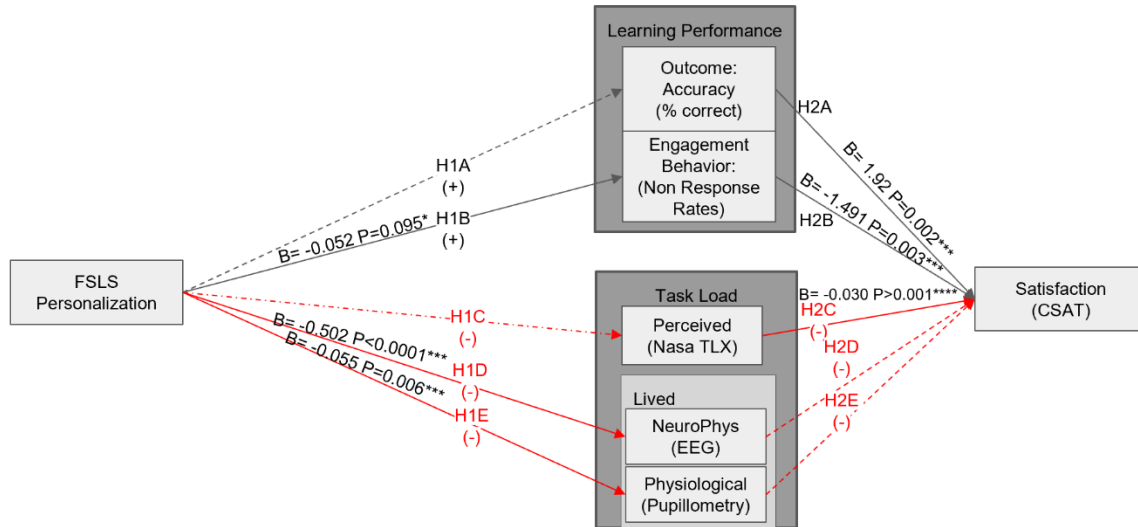


Note. Four indicators are presented across task order. Panels share the same x-axis to illustrate the temporal pattern across tasks (Task 1, Task 2, Task 3). Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is ***, $p < .001$ is ****.

5.8 Research Model Overview

The analysis described thus far has addressed each component of the proposed model by testing hypotheses H1A-H1E and H2A-H2E as summarized in Figure 7, which visualizes how each supported or unsupported hypothesis maps onto the research model (first presented in Figure 1).

Figure 7. Updated Research Model Incorporating Results from Hypothesis Testing



Note. Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is *** $p < .001$ is ****

In summary, personalization did not yield significant differences in learning accuracy (H1A) or in subjective task load, as measured by the NASA TLX (H1C). One behavioural indicator of engagement, the nonresponse rate, showed marginal effects in favour of personalization (H1B). Both physiological measures supported the expected effect of personalization on task load: EEG frontal theta-to-parietal alpha ratios (H1D) and pupil dilation (H1E) were significantly lower in the personalized condition while interacting with the LLM. Satisfaction outcomes were explained by several predictors. Accuracy (H2A), nonresponse rate (H2B), and NASA TLX scores (H2C) significantly contributed to satisfaction ratings, whereas EEG task load (H2D) and pupil size (H2E) did not.

6. Discussion

The present study examined whether persona-level adaptation of an AI mathematics tutor to learners' Felder-Silverman profiles would enhance learning effectiveness and user experience while reducing task load. Results revealed a complex pattern of behavioural activity: personalization significantly reduced physiological indices of task load, as measured by EEG and pupillometry, and marginally improved behavioural engagement (nonresponse rates), yet failed to produce detectable changes in learning accuracy or perceived mental effort. These findings contribute to the theoretical understanding of AI-mediated learning by demonstrating that a contested instructional design framework, when operationalized through system-prompt adaptation, can influence the cognitive processes underlying learning without necessarily improving immediate performance outcomes. Moreover, the divergence between physiological and perceptual task-load measures highlights methodological considerations essential for evaluating AI-mediated learning interactions.

6.1 Interpretation

The substantial reduction in EEG frontal-theta-to-parietal-alpha ratios under personalization effects persisted into post-test periods, demonstrating that the benefits extended beyond tutoring interactions to subsequent knowledge retrieval. Pupillometry data partially corroborated these findings: no change in mean pupil size during the post-test was observed; however, significantly smaller pupil diameters during the LLM interaction reflected reduced cognitive effort and arousal under personalized conditions (Krejtz et al., 2018; Rozado & Dunser, 2015). Pupil dilation decreased progressively across tasks, suggesting gradual adaptation and emerging

procedural fluency (Figure 6, Panel B). Despite weak correlations among individual task-load indices reflecting their distinct temporal resolutions and variance sources, the convergence of personalization outcomes between EEG and pupillometry supports the inference that persona-level adaptation reduced cognitive demands.

6.1.1 Understanding the Dissociation Between Task Load Indices

The correlations observed among our three task-load indices (perceived load, EEG, and pupillometry) align with established patterns in human-computer interaction and neuroscience research. Meta-analyses and empirical studies consistently report modest convergent validity ($r \approx .2-.5$) between physiological and self-report measures of cognitive load, with frequent task-dependent divergences (Ayres et al., 2021; Minkley et al., 2021; Rozado & Dunser, 2015). Several mechanisms explain these dissociations, each of which is relevant to the present findings.

One explanation lies in the mismatch of temporal dynamics and measurement granularity across modalities. Physiological indices such as EEG and pupillometry operate on rapid timescales that capture transient changes in effort and arousal, whereas self-reports integrate these fluctuations into a single retrospective judgment (Minkley et al., 2021). Pupillometry, recorded at 60 Hz, reflects variations in arousal and effort within 200-500 milliseconds of cognitive demand. EEG offers even greater temporal precision, with 16 frontal-parietal theta-alpha ratios, recorded at 500 hz, computed over six-second epochs, yielding a temporally rich index of task load that enhances sensitivity to brief neural adjustments (Ayres et al., 2021). In contrast, the NASA-TLX is administered once per block and requires participants to recall and average their impressions

across an entire seven-minute session. This mismatch inevitably limits convergence: momentary reductions in neural effort are often averaged out or distorted by memory biases such as recency effects or global effort heuristics (Ayres et al., 2021).

Beyond temporal resolution, dissociations may also stem from differences in metacognitive awareness. Learners do not have direct introspective access to their working-memory load or neural efficiency. Consequently, self-reports reflect perceived difficulty rather than actual cognitive demand (Hart & Staveland, 1988). This perception is mediated by metacognitive monitoring, a higher-order process influenced by factors such as task novelty, confidence, and prior expectations, which may operate independently of physiological effort (Hart & Staveland, 1988).

Physiological measures are further complicated by their sensitivity to affective and environmental factors. For instance, changes in pupil diameter and EEG power can be driven by emotional arousal, lighting, or stress rather than by task complexity alone (Krejtz et al., 2018; Rozado & Dunser, 2015). Although the study environment controlled for such variables, interacting with an AI tutor may have elicited mild anxiety or curiosity, affective states known to amplify physiological activation independently of cognitive load (Minkley et al., 2021). Lastly, Individual differences likely added further variability to these relationships. Variations in emotion regulation, metacognitive insight, and prior mathematics experience have been shown to moderate the correspondence between physiological and subjective indices (Ayres et al., 2021; Marek & Dosenbach, 2018).

Together, these factors clarify why physiological measures revealed benefits of personalization that participants did not consciously report. The observed dissociation does not invalidate either class of measurement; rather, it highlights their complementarity. Physiological indices capture the efficiency of cognitive processing, while self-reports capture the subjective experience of effort. Both perspectives are necessary for understanding how AI-mediated learning systems influence cognition, revealing that neurophysiological and introspective dimensions of task load, though only loosely correlated, provide distinct and equally valuable insights into learner experience.

6.1.2 Behavioural Indicators of Engagement: Nonresponse Rates and Formula Use

Although learning accuracy did not differ between conditions, nonresponse rates were lower, although marginally significant, in the treatment condition. This subtle behavioural shift provides insight into how personalization influenced participants' engagement with the assessment.

Nonresponse rates reflect not only knowledge gaps but also engagement and willingness to attempt a response (Anderson, 1982; Ericsson et al., 1993; Martin & Bolliger, 2018). By reducing nonresponse rates, personalization may have fostered a sense of competence or lowered the perceived stakes of guessing, encouraging participants to engage more fully with difficult items.

This interpretation is supported by the significant positive association between nonresponse rates and satisfaction. Participants who left fewer questions unanswered reported higher satisfaction with the AI tutor, suggesting that behavioural engagement mediated subjective experience. Personalization did not make the material easier, as accuracy remained unchanged, but it may

have made the learning process feel more manageable or aligned with participants' cognitive preferences, reducing avoidance behaviour (Bolliger & Martindale, 2004; Martin & Bolliger, 2018).

Complementing this pattern, transcript analyses revealed that personalization marginally increased the proportion of participant messages containing mathematical formulas. Participants in the treatment condition were more likely to engage with symbolic notation, either by copying formulas from the tutor's responses or by generating their own expressions during problem-solving. This behavioural trace suggests deeper interaction with the mathematical content, consistent with Cognitive Load Theory's emphasis on germane processing (Sweller, 2019). When learners actively manipulate symbols rather than passively reading explanations, they are more likely to construct robust mental schemas (Sweller, 2019).

However, participants in the treatment condition sent fewer messages. One interpretation is that personalized explanations were more efficient, requiring fewer clarification exchanges.

Alternatively, reduced message frequency could reflect satisfaction with initial responses, reducing the need for follow-up questions. Either way, the combination of lower nonresponse rates, higher formula use, and fewer messages paints a picture of more efficient engagement, in which participants in the treatment condition received adequate support with less communicative overhead, that is, elements of the text-based dialogue that are not essential to the learner's schema formation.

6.1.3 Absence of Immediate Learning Gains: Limitations and Explanations

The failure to detect differences in accuracy between conditions (H1A) requires careful interpretation. Several factors may have obscured learning effects. First, the mathematics assessments were challenging and designed to require conceptual understanding rather than procedural recall. Mean accuracy hovered near 38% across conditions, well below the ceiling level, meaning that task performance was far from the upper end (Over 90%) of the measurement scale and that the material presented substantial difficulty for participants (Gravetter & Forzano, 2018). This floor effect may have compressed variability, making it difficult to detect modest improvements attributable to personalization. If the assessments had included a wider range of difficulty or emphasized procedural fluency (for example, solving quadratic equations step by step rather than identifying vertex properties) and allowed more time, personalization effects might have emerged more clearly.

Second, the 7-minute tutoring sessions, though sufficient to elicit physiological differentiation, may have been too brief to produce measurable learning gains on complex topics such as logarithms and parabolas. Intelligent tutoring systems often require sustained interaction across multiple sessions before performance benefits manifest (Létourneau et al., 2025; Yi et al., 2024). In laboratory settings, single-session studies frequently observe process changes such as reduced cognitive load and increased engagement that precede outcome changes on more basic tasks (Liu et al., 2025; Lyu et al., 2025). Our results fit this pattern: personalization improved learning efficiency (lower task load and higher engagement) but did not produce measurable improvements in learning accuracy.

Third, the scope and abstract nature of the mathematical content may have exceeded both the system prompt's personalization logic and the generative model's ability to adjust its instructional approach. The FSLS-based adaptations, which distinguished between sequential and global or sensing and intuitive learners, operated at a pedagogical meta-level by adjusting explanatory style rather than tailoring content difficulty or prerequisite scaffolding. Participants lacking foundational knowledge, such as those unfamiliar with logarithms as inverse operations, may have struggled regardless of whether explanations emphasized concrete examples or abstract relationships during a brief seven-minute session. Future work could integrate adaptive content selection, including prerequisite diagnostic assessments, alongside persona adaptation to ensure that personalization addresses both how and what is taught.

Finally, individual differences in learning styles themselves remain contested. Although FSLS dimensions provided a structured framework for operationalizing personalization, psychometric critiques question whether matching instruction to self-reported preferences improves outcomes (Felder & Je, 2005; Graf et al., 2007; Pashler et al., 2008). Our results neither validate nor refute FSLS as a theoretical model; rather, they demonstrate that persona-level adaptation can influence cognitive processes such as task load and engagement even when grounded in a debated framework. The value of personalization may lie less in fidelity to any specific learning adaptation taxonomy and more in its capacity to vary instructional discourse, thereby increasing the likelihood that at least some learners encounter a format congruent with their processing preferences (Liu et al., 2025; Samuel et al., 2024; Zheng et al., 2024).

When comparing learning-style conditions, Group A (sensing/intuitive: concrete versus abstract content framing) displayed a higher average EEG task load than Group B (sequential/global: linear versus holistic sequencing). Although this difference is not statistically significant, the trend suggests that the explanation exerts stronger effects than the sequencing. Concrete examples may reduce representational friction more effectively by providing robust cognitive anchors for abstract mathematics (Graf et al., 2007; Samia & Amirat, 2017). However, neither learning-style adaptation type improved learning outcomes nor satisfaction, reinforcing that personalization influenced process efficiency rather than outcome effectiveness within brief exposures. Task-order analyses across the three learning blocks revealed systematic decreases in nonresponse rates and physiological task load, as measured by pupillometry, across the first two successive blocks, reflecting meta-strategic learning and adaptation to the interface and assessment format rather than content mastery, as accuracy remained stable (Figure 7). While pupillometry showed a decrease in task load across tasks, EEG detection did not yield any significant changes. It may be interpreted from the pupillometry and behavioural results that the within-subject design introduced carryover effects whereby procedural strategies and interface familiarity transferred across blocks.

6.1.4 Satisfaction and User Experience

Satisfaction ratings provided insight into the experiential correlates of learning and task load. Accuracy positively predicted satisfaction, as did lower nonresponse rates, confirming that participants derived satisfaction from successful performance and active engagement rather than from interface features alone. Perceived task load strongly predicted dissatisfaction, consistent

with Cognitive Load Theory's emphasis on reducing extraneous burden (Hart & Staveland, 1988; Sweller, 2019).

Critically, neither EEG task load nor pupil dilation significantly predicted satisfaction ($p = .804$ and $p > .05$, respectively). This null result further reinforces the dissociation between physiologically measured efficiency and the subjective experience of mental effort, as addressed in an earlier section. Participants did not consciously register the neural or arousal reductions produced by personalization; thus, these efficiency gains did not enhance their reported satisfaction. This pattern has important implications for the design of neuroadaptive systems. If physiological improvements remain subliminal, they may not drive user acceptance or perceived value unless made salient through feedback or framing. For example, a neuroadaptive tutor could display brief, non-evaluative indicators that the learner is maintaining an efficient level of cognitive effort, allowing users to connect these underlying changes to their ongoing experience.

The manipulation check, which consisted of a single item asking whether the tutor tailored explanations to individual learning needs, revealed no significant difference between conditions ($p > .10$). This finding indicates that participants did not consciously detect personalization. The absence of conscious detection could reflect insufficient salience of the persona adaptations, as participants may have attributed stylistic differences to randomness or topic variation. In any case, the absence of conscious detection combined with robust physiological effects suggests that implicit personalization, or adaptation that operates below conscious awareness, can nonetheless influence cognitive processing (Minkley et al., 2021).

6.2 Methodological Contributions: Multimodal Measurement

This study advances AI-education research methodology by integrating EEG and pupillometry with self-report and behavioural measures within a single design. Most AI tutoring evaluations rely solely on performance metrics such as accuracy or satisfaction surveys, leaving the cognitive processes underlying learning opaque (Gkintoni et al., 2025; Sarailoo et al., 2022). By quantifying task load via neurophysiological measures, we revealed efficiency gains that are invisible to traditional assessments. This approach is particularly relevant for neuroadaptive systems, which adjust in real time based on brain states and require validated physiological markers of cognitive demand (Beauchemin et al., 2024; Zander & Kothe, 2011).

Our use of 16 frontal-parietal EEG ratios rather than single-electrode indices enhanced robustness by averaging across spatial variability and reducing susceptibility to localized artifacts (Sollazzo et al., 2025). Six-second epochs balanced temporal resolution with signal stability, capturing fluctuations in working memory engagement without excessive noise. Pupillometry complemented EEG by providing a second physiological channel sensitive to arousal and effort, enabling cross-validation. The weak inter-method correlations, rather than undermining validity, reflect the multidimensional nature of cognitive load as a construct that encompasses neural efficiency, autonomic arousal, and subjective experience (Ayres et al., 2021). Multimodal measurement acknowledges this complexity and yields richer insights than any single method.

6.3 Practical Implications for AI-Mediated Education

Despite the absence of reported positive effects on immediate accuracy, the demonstrated reductions in physiological task load have practical significance for scalable educational technologies. Neuroadaptive systems that adjust content presentation in real time based on EEG or pupillometric feedback have the potential to sustain learners within optimal engagement zones, preventing both cognitive overload and disengagement (Zander & Kothe, 2011). Our findings suggest that such systems need not achieve perfect personalization accuracy, even coarse adaptations based on debated frameworks can reduce neural effort.

For educators and instructional designers, the results suggest that efficiency and effectiveness are dissociable learning outcomes. An AI tutor that lowers cognitive load may not immediately boost test scores but could enhance long-term retention, reduce frustration, or increase willingness to persist through challenging material, outcomes not captured by single-session performance metrics. Evaluations of AI tutoring systems should therefore employ multimodal assessments that integrate physiological, behavioural, and self-report measures across extended time horizons to capture the full spectrum of benefits.

For policymakers and edtech developers, the study highlights the feasibility of persona-level personalization through system-prompt engineering, a low-cost, scalable approach that requires no changes to visual interface design or multimedia content. Large language models such as Mistral, GPT-4, or Claude can implement adaptive personas with minimal infrastructure, democratizing access to personalized tutoring. However, the absence of conscious detection and the modest behavioural effects caution against overstating such adaptations as pedagogical

solutions. Personalization is most likely to yield measurable benefits when combined with adaptive content selection, scaffolded practice, and formative feedback, which are already integrated into tutoring modes in popular LLMs like ChatGPT (Chatterji et al., 2025; OpenAI, 2025).

6.4 Limitations and Future Directions

Several constraints beyond those embedded in the interpretations above warrant consideration. The university-educated adult sample limits generalizability to K-12 students or individuals with lower educational attainment, populations that may exhibit different responses. Laboratory constraints imposed by EEG and eye-tracking equipment (e.g., stillness requirements and a no-calculator policy) may have increased intrinsic and extraneous cognitive demands unrelated to tutor design.

Future research should test whether gains in neurophysiological efficiency predict longitudinal outcomes using multi-session designs conducted over a longer period. Pre-tests would enable computation of learning-gain scores and clarify where in the learning pipeline personalization exerts the strongest effects. The weak correlations among task-load measures (EEG, pupillometry, and NASA-TLX) highlight the need for time-locked micro-assessments.

Embedding brief Likert-scale probes within tutoring sessions (for example, every 60 seconds: “Rate your current mental effort, 1-5”) would align self-reports with granular physiological data, enabling lagged correlation analyses to estimate subjective awareness latency. Lastly, despite its widespread use, the FSLs instrument lacks strong psychometric validation (Graf et al., 2007).

Future work could employ alternative frameworks, such as cognitive preferences derived from

working-memory capacity or aptitude-treatment interaction designs, to determine whether personalization effects depend on stable individual differences or transient task states (Lyu et al., 2025; Zobel & Meinel, 2025).

7. Conclusion

This study examined whether adapting an AI mathematics tutor's instructional persona to learners' Felder-Silverman profiles enhances learning effectiveness, user experience, and cognitive efficiency. Personalization produced significant reductions in physiological task load, as indexed by EEG frontal theta to parietal alpha ratios and pupil dilation, indicating improved processing efficiency. Behaviourally, participants in the personalized condition showed lower nonresponse rates and a higher proportion of math-formula entries, suggesting greater engagement. Satisfaction analyses revealed that participants were most satisfied when performance was high, and effort felt manageable: accuracy positively predicted satisfaction, whereas higher nonresponse and perceived task load predicted dissatisfaction. However, neither EEG nor pupil-based measures of task load predicted satisfaction, suggesting that physiological efficiency gains were not consciously perceived. Consequently, participants may have benefited cognitively without registering those benefits affectively or attributing them to personalization.

These results refine debates on the educational value of learning-style frameworks. Although the Felder-Silverman model lacks strong psychometric grounding, implementing it through system-prompt adaptation measurably affected cognitive efficiency. Personalization, therefore, appeared to lower information-processing barriers rather than directly improve performance. Cognitive

Load Theory supports this interpretation, as reducing extraneous load frees resources for germane processing while maintaining type 2 activation, consistent with Dual Process learning theory (Evans & Stanovich, 2013; Sweller et al., 2011). In this study, efficiency gains preceded measurable improvements in performance or satisfaction.

The divergence between physiological and self-reported task load was evidenced by weak correlations among EEG, pupil, and NASA TLX measures. The lack of subjective awareness of reduced neural effort highlights that efficiency gains can occur below conscious perception yet still shape long-term retention and persistence. This dissociation has methodological and practical implications. For BCI researchers developing neuroadaptive educational technologies, validated physiological markers provide a foundation for neurofeedback even when learners cannot introspect their cognitive states. For researchers in education more broadly, the findings emphasize the need to triangulate performance, self-report, and physiological indicators to capture the multifaceted effects of AI-mediated learning.

Despite acknowledged design and sampling constraints, the study advances understanding of AI-mediated learning. It demonstrates that persona-level adaptation, achieved solely through prompt engineering, can reduce cognitive load without altering interfaces (Lyu et al., 2025; Samuel et al., 2024; Zobel & Meinel, 2025). It also illustrates how combining physiological and perceptual data reveals hidden efficiency benefits that conventional metrics overlook. Finally, it shows that even theoretically disputed frameworks can yield practical gains when operationalized as design heuristics (Zheng et al., 2024).

Future research should test real-time neuroadaptive tutors that monitor EEG or pupillometric indicators and dynamically adjust instruction to keep cognitive load within an optimal threshold. Integrating persona-level adaptation with adaptive content selection and scaffolded feedback could reveal synergistic effects across personalization layers.

The global learning crisis, characterized by declining proficiency despite increased investment (OECD, 2023b; UNESCO, 2023), demands innovative solutions that enhance both access and instructional effectiveness. AI tutoring systems, empowered by advances in large language models and brain-sensing technologies, hold promise for democratizing high-quality personalized education at scale (Xia et al., 2024). However, realizing this potential requires rigorous empirical evaluation grounded in cognitive science, careful attention to both process and outcome measures, and acknowledgment that efficiency gains, although less visible than accuracy improvements, may confer long-term advantages for learners navigating increasingly complex knowledge domains (Krol & Zander, 2017; Sweller, 2019). By quantifying the cognitive cost of learning alongside its outcomes, research of this kind lays the groundwork for next-generation educational technologies capable of sustaining engagement, optimizing mental effort, and ultimately improving learning trajectories for diverse populations.

References

- Alrashidi, O., Phan, H. P., & Ngu, B. H. (2016). Academic Engagement: An Overview of Its Definitions, Dimensions, and Major Conceptualisations. *International Education Studies*, 9(12), 41. <https://doi.org/10.5539/ies.v9n12p41>
- Anderson, J. R. (1982). Acquisition of Cognitive Skill. *Psychological Review*, 89(4), 369-406. <https://doi.org/10.1037/0033-295X.89.4.369>
- Antonenko, P. P., Paas, F., Grabner, R., & Gog, T. (2010). Using Electroencephalography to Measure Cognitive Load. *Educational Psychology Review*, 22, 425-438. <https://doi.org/10.1007/s10648-010-9130-y>
- Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load [Review]. *Frontiers in Psychology, Volume 12 - 2021*. <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.702538>
- Beauchemin, N., Charland, P., Karran, A., Boasen, J., Tadson, B., Sénécal, S., & Léger, P.-M. (2024). Enhancing learning experiences: EEG-based passive BCI system adapts learning speed to cognitive load in real-time, with motivation as catalyst. *Frontiers in Human Neuroscience*, 18, 1416683. <https://doi.org/10.3389/fnhum.2024.1416683>
- Blumenfeld, P., & Paris, A. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research - REV EDUC RES*, 74, 59-109. <https://doi.org/10.3102/00346543074001059>
- Bolliger, D., & Martindale, T. (2004). Key factors for determining student satisfaction in online courses. 3, 61-67.
- Capotosto, P., Babiloni, C., Romani, G. L., & Corbetta, M. (2009). Frontoparietal Cortex Controls Spatial Attention through Modulation of Anticipatory Alpha Rhythms. *The Journal of Neuroscience*, 29(18), 5863-5872. <https://doi.org/10.1523/jneurosci.0539-09.2009>
- Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, 18(8), 414-421. <https://doi.org/https://doi.org/10.1016/j.tics.2014.04.012>
- Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How People Use ChatGPT. *National Bureau of Economic Research Working Paper Series, No. 34255*. <https://doi.org/10.3386/w34255>

- Cohen, J. (1992). A power primer [doi:10.1037/0033-2909.112.1.155]. *112*, 155-159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201-215. <https://doi.org/10.1038/nrn755>
- Ericsson, K., Krampe, R., & Tesch-Roemer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review*, *100*, 363-406. <https://doi.org/10.1037//0033-295X.100.3.363>
- Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223-241. <https://doi.org/10.1177/1745691612460685>
- Farris, P. W., Bendle, N., Pfeifer, P., & Reibstein, D. (2010). *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Pearson Education.
- Felder, R. (1988). Learning and Teaching Styles in Engineering Education. *Journal of Engineering Education*, *78*, 674-681.
- Felder, R., & Je, S. (2005). Applications, reliability and validity of the Index of Learning Styles. *International Journal of Engineering Education*, *21*, 103-112.
- Fredette, M. C., Shang Lin; Tech3Lab Research Team. (2025). *UX data analysis toolkit: A practical guide for hypothesis testing using SAS (Version 1.0, Summer 2025)* (version 1.0). Tech3Lab – HEC Montréal.
- Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy. *Brain Sci*, *15*(2). <https://doi.org/10.3390/brainsci15020203>
- Graf, S., Viola, S. R., Leo, T., & Kinshuk. (2007). In-Depth Analysis of the Felder-Silverman Learning Style Dimensions. *Journal of Research on Technology in Education*, *40*(1), 79-93. <https://doi.org/10.1080/15391523.2007.10782498>
- Graf, S., Lin, T., & Kinshuk. (2008). The relationship between learning styles and cognitive traits – Getting additional information for improving student modelling. *Computers in Human Behavior*, *24*(2), 122-137. <https://doi.org/https://doi.org/10.1016/j.chb.2007.01.004>
- Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(267), 1-13. <https://doi.org/10.3389/fnins.2013.00267>

- Gravetter, F. J., & Forzano, L. A. B. (2018). *Research Methods for the Behavioural Sciences* (6 ed.). Cengage Learning.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139-183). North-Holland.
[https://doi.org/https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/https://doi.org/10.1016/S0166-4115(08)62386-9)
- Haug, J., Fischer, D., & Hagel, G. (2023). Development of a short form of the index of learning styles for the use in adaptive learning systems. Proceedings of the 5th European Conference on Software Engineering Education,
- Healy, G. F., Boran, L., & Smeaton, A. F. (2015). Neural Patterns of the Implicit Association Test [Original Research]. *Frontiers in Human Neuroscience, Volume 9 - 2015*.
<https://doi.org/10.3389/fnhum.2015.00605>
- Institute, S. A. S. (2024). *SAS OnDemand for Academics*. In SAS Institute Inc.
<https://www.sas.com/ondemand>
- Jasper, H. H. (1958). The 10-20 electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology, 10*, 371-375
- Karmakar, S., Kamilya, S., Dey, P., Guhathakurta, P. K., Dalui, M., Bera, T. K., Halder, S., Koley, C., Pal, T., & Basu, A. (2023). Real time detection of cognitive load using fNIRS: A deep learning approach. *Biomedical Signal Processing and Control, 80*, 104227.
<https://doi.org/https://doi.org/10.1016/j.bspc.2022.104227>
- Karran, A. J., Charland, P., Trempe-Martineau, J., Ortiz de Guinea Lopez de Arana, A., Lesage, A. M., Sénécal, S., & Léger, P. M. (2025). Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education. *NPJ Sci Learn, 10*(1), 44.
<https://doi.org/10.1038/s41539-025-00333-2>
- Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews, 29*(2), 169-195.
[https://doi.org/https://doi.org/10.1016/S0165-0173\(98\)00056-3](https://doi.org/https://doi.org/10.1016/S0165-0173(98)00056-3)
- Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences, 16*(12), 606-617.
<https://doi.org/https://doi.org/10.1016/j.tics.2012.10.007>
- Kolb, D. (1984). *Experiential Learning: Experience As The Source Of Learning And Development* (Vol. 1). <http://academic.regis.edu/ed205/Kolb.pdf>

- Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE*, 13(9), e0203629. <https://doi.org/10.1371/journal.pone.0203629>
- Krol, L. R., & Zander, T. O. (2017). Passive BCI-based Neuroadaptive Systems. Graz Brain-Computer Interface Conference, <https://doi.org/10.3217/978-3-85125-533-1-46>
- Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10(1), 29. <https://doi.org/10.1038/s41539-025-00320-7>
- Liu, B., Zhang, J., Lin, F., Jia, X., & Peng, M. (2025). One size doesn't fit all: A personalized conversational tutoring agent for mathematics instruction. Companion Proceedings of the ACM on Web Conference 2025, <https://doi.org/10.1145/3701716.3717527>
- Lyu, B., Li, C., Li, H., Oh, H., Song, Y., Zhu, W., & Xing, W. (2025). *Exploring the Role of Teachable AI Agents' Personality Traits in Shaping Student Interaction and Learning in Mathematics Education* Proceedings of the 15th International Learning Analytics and Knowledge Conference, <https://doi.org/10.1145/3706468.3706532>
- Marek, S., & Dosenbach, N. U. F. (2018). The frontoparietal network: function, electrophysiology, and importance of individual precision mapping. *Dialogues in Clinical Neuroscience*, 20(2), 133-140. <https://doi.org/10.31887/dcns.2018.20.2/smarek>
- Martin, F., & Bolliger, D. U. (2018). Engagement Matters: Student Perceptions on the Importance of Engagement Strategies in the Online Learning Environment. *Online Learning*, 22(1). <https://doi.org/10.24059/olj.v22i1.1092>
- May, D. K. (2009). *Mathematics Self-Efficacy and Anxiety Questionnaire* [Doctoral dissertation, The University of Georgia]. Athens, Georgia.
- Minkley, N., Xu, K., & Krell, M. (2021). Analyzing Relationships Between Causal and Assessment Factors of Cognitive Load: Associations Between Objective and Subjective Measures of Cognitive Load, Stress, Interest, and Self-Concept. *Frontiers in Education*, 6. <https://doi.org/10.3389/educ.2021.632907>
- Mistral, A. I. t. (2024, 2024/07/24). Large Enough: Mistral Large 2 announced. <https://mistral.ai/news/mistral-large-2407>
- OECD. (2023a). *PISA 2022 results (Volume I and II) - Country notes: Canada*. OECD Publishing. Retrieved September, 2025 from https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/canada_901942bb-en.html

- OECD. (2023b). *PISA 2022 results (Volume I and II) - Country notes: United States*. OECD Publishing. Retrieved Sept 2025 from https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes_ed6fbcc5-en/united-states_a78ba65a-en.html
- OpenAI. (2025, July 29, 2025). *Introducing study mode*. openAI. Retrieved Sept, 2025 from <https://openai.com/index/chatgpt-study-mode>
- Paivio, A. (1990). *Mental Representations: A dual coding approach*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195066661.001.0001>
- Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE*, 19(5), e0304013. <https://doi.org/10.1371/journal.pone.0304013>
- Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest*, 9(3), 105-119. <https://doi.org/10.1111/j.1539-6053.2009.01038.x>
- Python Software, F., & Van Rossum, G. (2022). *Python (Version 3.11) [Computer software]*. In Python Software Foundation. <https://www.python.org/>
- Rozado, D., & Dunser, A. (2015). Combining EEG with Pupillometry to Improve Cognitive Workload Detection. *Computer*, 48(10). <https://doi.org/10.1109/MC.2015.314>
- Samia, D., & Amirat, A. (2017). Adaptation with Four Dimensional Personalization Criteria Based on Felder Silverman Model. *International Journal of Distance Education Technologies*, 15(4), 1-20. <https://doi.org/10.4018/IJDET.2017100101>
- Samuel, V., Zou, H. P., Zhou, Y., Chaudhari, S., Kalyan, A., Rajpurohit, T., Deshpande, A., Narasimhan, K., & Murahari, V. (2024). Personagym: Evaluating persona agents and llms. <https://doi.org/10.48550/arXiv.2407.18416>
- Sarailoo, R., Latifzadeh, K., Amiri, S. H., Bosaghzadeh, A., & Ebrahimpour, R. (2022). Assessment of instantaneous cognitive load imposed by educational multimedia using electroencephalography signals [Original Research]. *Frontiers in Neuroscience, Volume 16 - 2022*. <https://doi.org/10.3389/fnins.2022.744737>
- Sollazzo, K., Karran, A. J., Rolon-Merette, T., Stefanescu, I. M., Coursaris, C., Léger, P.-M., & Sénécal, S. (2025, 2025/12). *Pathways to Optimal Learning: Task Load-Driven Neuroadaptive Adaptation and Motivational Incentives* NeuroIS Retreat 2025,
- Spitzer, M. W. H., & Moeller, K. (2024). Performance increases in mathematics within an intelligent tutoring system during COVID-19 related school closures: a large-scale longitudinal evaluation. *Computers and Education Open*, 6, 100162. <https://doi.org/https://doi.org/10.1016/j.cao.2024.100162>

- Sweller, J. (2019). Cognitive load theory and educational technology. *Educational Technology Research and Development*, 68(1), 1-16. <https://doi.org/10.1007/s11423-019-09701-3>
- Sweller, J., Ayres, P. L., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. <https://doi.org/10.1007/978-1-4419-8126-4>
- UNESCO. (2023). *Highlights from the global report on teachers: What you need to know*. <https://www.unesco.org/en/articles/highlights-global-report-teachers-what-you-need-know>
- Walkington, C., & Bernacki, M. L. (2019). Personalizing Algebra to Students' Individual Interests in an Intelligent Tutoring System: Moderators of Impact. *International Journal of Artificial Intelligence in Education*, 29(1), 58-88. <https://doi.org/10.1007/s40593-018-0168-1>
- Wang, B., Rau, P.-L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, 42(9), 1324-1337. <https://doi.org/10.1080/0144929X.2022.2072768>
- Wascher, E., Rasch, B., Sanger, J., Hoffmann, S., Schneider, D., Rinkenauer, G., & Heuer, H. (2014). Frontal theta activity reflects distinct aspects of mental fatigue. *Biological psychology*, 96, 57-65. <https://doi.org/https://doi.org/10.1016/j.biopsycho.2013.11.010>
- Xia, Q., Chiu, T. K. F., & Li, X. (2024). A scoping review of BCIs for learning regulation in mainstream educational contexts. *Behaviour & Information Technology*, 43(10), 2096-2117. <https://doi.org/10.1080/0144929X.2023.2241559>
- Yi, L., Liu, D., Jiang, T., & Xian, Y. (2024). The Effectiveness of AI on K-12 Students' Math Learning: A Systematic Review and Meta-Analysis. *International Journal of Science and Mathematics Education*, 1-22. <https://doi.org/10.1007/s10763-024-10499-7>
- Zander, T. O., & Kothe, C. (2011). Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *J Neural Eng*, 8(2), 025005. <https://doi.org/10.1088/1741-2560/8/2/025005>
- Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2024). When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. Findings of the Association for Computational Linguistics: EMNLP 2024, <https://doi.org/10.18653/v1/2024.findings-emnlp.888>
- Zobel, T., Meinel, C. (2025). Chatbot Personas as a Gateway to Enhanced Learning Experiences. In: Arai, K. (eds) *Advances in Information and Communication. FICC 2025. Lecture Notes in Networks and Systems*, vol 1283. Springer, Cham. https://doi.org/10.1007/978-3-031-84457-7_13

Chapter 3: Conclusion

This thesis examined how persona-level personalization in an AI tutor influences the user's experience in terms of the cognitive and behavioural processes that support learning. The study integrated electroencephalography, pupillometry, and interaction-level behavioural measures to determine whether aligning a math AI tutor's persona with learners' Felder-Silverman learning style improves learning outcomes. Across these measures, the findings suggest that personalization may have an effect on the learner's cognitive load during the task of engaging with an AI math tutor, even when immediate performance outcomes remain unchanged.

Learners who interacted with the personalized tutor showed lower neurophysiological task load. This effect was evident in reduced frontal theta to parietal alpha task load ratios and smaller pupil dilations during tutoring. These patterns indicate more efficient management of cognitive effort when instructional delivery aligns with a learner's preferred information-processing mode. Behaviourally, participants in the personalized condition responded more consistently to quiz items and used symbolic notation more often during the interactions with the tutor during the learning phase, suggesting greater engagement, reduced avoidance, and deeper involvement in mathematical procedures (Ericsson et al., 1993; Martin & Bolliger, 2018). Accuracy and satisfaction remained stable across conditions, consistent with evidence that conceptual mathematics often requires extended periods of practice before improvements can be reliably measured. Together, these results show that personalization influences the quality of learners' cognitive engagement even when short-term learning gains do not shift.

Theoretical Contributions

The findings contribute to the theoretical foundations introduced in Chapter 1. Cognitive Load Theory (CLT) proposes that instructional misalignment creates unnecessary demands on working memory (Sweller, 2019). The reductions in physiological task load observed here support the mechanisms discussed in CLT and illustrate how minimal changes in a tutor's communicative persona can reduce extraneous demands on learners (Sweller, 2019). The results also align with Dual Process Theory, in which personalized instruction appears to reduce the need for deliberate, controlled processing (Type 2) relative to the higher cognitive load in the control condition (Evans & Stanovich, 2013). These complementary perspectives indicate that persona personalization may support a more effective distribution of cognitive effort during the tutoring interactions. The reduced task load observed in the study reflects this possibility. This interpretation does not assume that learning style models predict performance. Instead, the study treats learning style categories as practical labels for describing interactional preferences that can shape how learners process explanations moment by moment.

Practical Implications

The study also offers insights for the design of AI-mediated learning systems. The combination of cognitive efficiency gains and increased engagement suggests that persona-based adaptation can support learners who might otherwise disengage when explanations feel unfamiliar or cognitively taxing. Educational institutions considering the integration of AI tutors may benefit from personalization features that tailor not only content difficulty, but also communication style. At the same time, scalable deployment introduces several ethical considerations. Clear guidelines for how tutors should be designed and used, safeguards against over-reliance, and

transparent communication about model behaviour and user data requirements remain necessary to ensure that AI tools strengthen learning outcomes while minimizing externalities (Karran et al., 2025).

Future Research Directions

Several research avenues arise from this work. First, extending the duration of tutoring interactions would help determine whether reductions in cognitive effort eventually translate into improved accuracy, retention, and schema formation. Multi-session formats would also allow examination of whether cognitive efficiency accumulates over repeated exposures.

Second, future studies should broaden the personalization mechanisms tested. Adjusting difficulty, timing, and scaffolding based on real-time indicators of mental effort could reveal synergies between persona-level adaptation and other forms of educational tailoring.

Future research should refine the measurement of task load by improving both physiological and subjective assessment methods. In this study, correlations among EEG, pupillometry, and NASA TLX scores were modest, which suggests that cognitive load fluctuates over short timescales and that global post-session questionnaires are not sufficiently aligned with these rapid dynamics.

Embedding brief, time-locked micro assessments within the tutoring exchange, such as periodic single-item ratings of perceived mental effort, would allow researchers to synchronize subjective reports with specific neural events. When paired with event-dependent stimuli that define clear temporal anchors, this approach would enable more precise comparisons of fixed instructional moments and support lagged analyses that estimate the temporal relationship between subjective awareness and physiological markers of load. Together, these refinements would strengthen

multimodal interpretations of task load and improve the validity of inferences about cognitive processing in AI-mediated learning.

Lastly, ecological validity and generalizability should be enhanced through classroom-embedded research, a larger sample size, and a wider age range. Bringing AI tutors into authentic learning environments would clarify how personalization interacts with classroom dynamics, curriculum pacing, and learners' ongoing exposure to mathematics instruction across different ages.

Concluding Remarks

This thesis provides evidence that personalizing an AI mathematics tutor's communicative style through personas, via system-level prompts, influences how learners manage cognitive effort during learning. In a context where educational systems face increasing pressures and students have increased their usage of AI tools, understanding the cognitive implications of these systems is essential. The results show that even personalization based on a disputed framework, applied at the level of the tutor's persona in an accessible manner, can support more efficient cognitive processing and encourage sustained engagement. By integrating behavioural, subjective, and physiological evidence, this work contributes to a growing understanding of how AI-mediated instruction can be designed to support learners effectively and equitably. Continued research that integrates cognitive theory, adaptive technologies, and real-world learning contexts will be crucial for shaping responsible, evidence-based applications of AI in education.

References

- Ericsson, K., Krampe, R., & Tesch-Roemer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review*, 100, 363-406. <https://doi.org/10.1037//0033-295X.100.3.363>
- Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, 8(3), 223-241. <https://doi.org/10.1177/1745691612460685>
- Karran, A. J., Charland, P., Trempe-Martineau, J., Ortiz de Guinea Lopez de Arana, A., Lesage, A. M., Sénécal, S., & Léger, P. M. (2025). Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education. *NPJ Sci Learn*, 10(1), 44. <https://doi.org/10.1038/s41539-025-00333-2>
- Martin, F., & Bolliger, D. U. (2018). Engagement Matters: Student Perceptions on the Importance of Engagement Strategies in the Online Learning Environment. *Online Learning*, 22(1). <https://doi.org/10.24059/olj.v22i1.1092>
- Sweller, J. (2019). Cognitive load theory and educational technology. *Educational Technology Research and Development*, 68(1), 1-16. <https://doi.org/10.1007/s11423-019-09701-3>

Appendices

Appendix A: Participant Exclusion Criteria

Table 10 (Appendix A1). *Inclusion and Exclusion Criteria for Test Participants*

Inclusion/ Exclusion Criteria
Must be over 18 years of age;
Be able to speak and write French OR English;
No skin allergies or particular sensitivities;
No laser vision correction;
No neurological or psychiatric diagnosis;
Do not need glasses for computer work (contact lenses are permitted);
Have no astigmatism;
Do not have a pacemaker;
Does not suffer from epilepsy
Not come to the session intoxicated or under the influence of alcohol, marijuana or other recreational drugs;
Do not have a hairstyle that restricts access to the scalp.

Note. Participants must meet all the above criteria to participate in the study.

Appendix B: Post-Task Math Assessments

To access the full list of questions and answers used to assess math learning performance in all three subjects, follow this link: [V1_Maths_Questions_with_answers.pdf](https://hecmtl.sharepoint.com/:b:/r/sites/Tech3Lab/Documents%20partages/Projets/434%20-%20Mogwai/2.%20Collecte/V1_Maths_Questions_with_answers.pdf?csf=1&web=1&e=6r8hSp)
(https://hecmtl.sharepoint.com/:b:/r/sites/Tech3Lab/Documents%20partages/Projets/434%20-%20Mogwai/2.%20Collecte/V1_Maths_Questions_with_answers.pdf?csf=1&web=1&e=6r8hSp)

Appendix C: System-Level Persona Prompts

Follow the link below to access an Excel sheet which you can download containing three sheets.

1. Core System Level prompt: used in all conditions (control and treatment)
2. Learning Styles: Applied in the treatment condition for each Fielder Silverman learning style.
3. Task-based instructions: Specifications for each math subject which includes a starting message and configurations for optimal equation displaying.

Link to Document: [EN-System Prompts.xlsx](https://hecmtl.sharepoint.com/:x:/r/sites/Tech3Lab/Documents%20partages/Projets/434%20-%20Mogwai/2.%20Collecte/EN-System_Prompts.xlsx?d=w2dc3b89807a5451fb326d5cc6494ef03&csf=1&web=1&e=FZwbiV)

(https://hecmtl.sharepoint.com/:x:/r/sites/Tech3Lab/Documents%20partages/Projets/434%20-%20Mogwai/2.%20Collecte/EN-System_Prompts.xlsx?d=w2dc3b89807a5451fb326d5cc6494ef03&csf=1&web=1&e=FZwbiV)

Appendix D: Expanded Results Section

Table 11 (Appendix D1). *Descriptive Statistics table for Felder Silverman Learner Subgroups (Group A vs. Group B)*

Variables	Group A (<i>M</i>)	Group A (<i>SD</i>)	Group B (<i>M</i>)	Group B (<i>SD</i>)
TLX	51.05	26.27	50.23	22.68
EEG TL study	2.37	2.84	2.05	1.88
EEG TL Test	2.17	2.48	2.18	2.48
Pupillometry TL	0.79	0.38	0.82	0.30
Accuracy (Correct)	0.36	0.21	0.39	0.17
CSAT	4.20	1.38	4.58	1.39
Manipulation check	4.27	1.75	4.52	1.44
Nonresponse rate	0.30	0.26	0.29	0.23
Prompts with Formulas	0.02	0.07	0.02	0.06

Note. No significant values were found from the mixed linear regression; therefore, no inferential statistics were included.

Table 12 (Appendix D2). *Repeated Measures ANOVA Results for Task Order Comparison*

Variables	Task 1 (<i>M</i>)	Task 1 (<i>SD</i>)	Task 2 (<i>M</i>)	Task 2 (<i>SD</i>)	Task 3 (<i>M</i>)	Task 3 (<i>SD</i>)	<i>df</i>	<i>F</i>	Global <i>p</i>
Accuracy (Correct)	0.34	0.18	0.39	0.19	0.37	0.21	78	3.24	0.325
Nonresponse	0.35	0.25	0.27	0.26	0.27	0.23	74.71	2.89	0.061*
TLX	47.23	24.50	52.20	24.65	52.50	24.42	78	1.39	0.142
EEG TL Study	2.16	2.05	2.26	2.58	2.23	2.68	70	0.35	0.552
EEG TL Test	1.90	2.11	2.26	2.55	2.36	2.74	70	4.03	0.008**
Pupillometry TL	0.85	0.35	0.79	0.33	0.77	0.34	72	8.32	0.000****
Prompts with formula	0.02	0.05	0.04	0.09	0.01	0.04	78	2.42	0.031**
CSAT	4.45	1.43	4.35	1.45	4.40	1.50	78	0.14	0.835

Note. This table does not include pairwise analysis results; P-values represent global regression values. Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is ***, $p < .001$ is ****

Table 13 (Appendix D13). *Results from Repeated Measures ANOVA Comparing Control and Groups A and B*

Variable	Control (<i>M</i>)	Control (<i>SD</i>)	Group A (<i>M</i>)	Group A (<i>SD</i>)	Group B (<i>M</i>)	Group B (<i>SD</i>)	<i>DF</i>	<i>F</i>	<i>p</i>
TLX	48.80	24.61	55.15	26.99	49.79	21.60	99.65	0.64	0.556
EEG study	2.69	2.94	1.97	1.51	1.54	1.40	71.17	15.59	0.006**
EEG test	2.65	2.96	1.92	1.51	1.51	1.31	70.00	4.03	0.009**
pupillometry	0.82	0.84	0.77	0.36	0.80	0.80	74.82	4.03	0.095*
correct	0.38	0.19	0.35	0.22	0.40	0.17	100.30	0.65	0.556
CSAT	4.36	1.35	4.20	1.49	4.65	1.38	99.35	2.01	0.891
Manipulation	4.37	1.52	4.33	1.83	4.48	1.57	102.00	0.51	0.347
Nonresponse	0.31	0.26	0.29	0.24	0.28	0.23	93.08	0.97	0.490
Prompt with formula	0.01	0.04	0.03	0.09	0.03	0.08	74.82	4.03	0.095*

Note. This table does not include pairwise analysis results; P-values represent global regression values. Asterisks indicate significance levels $p < .10$ is *, $p < .05$ is **, $p < .01$ is ***, $p < .001$ is ****

Appendix E: Generative AI Use Disclosure Table

Table 14 (Appendix E1). *AGI Use Disclosure Table*

Stage in the process	Use of generative AI
Research Question	[0% AI] Research questions and problem formulation were developed independently based on the research literature and discussions with supervisors.
Literature review	[15% AI] Generative AI was used to summarize selected scholarly articles, extract conceptual themes, and support the initial structuring of the literature review. All summaries were checked against primary sources and integrated manually.
Experimental design	[20% AI] Generative AI supported the development and refinement of the system level prompts used to instantiate the personalized and non personalized tutor personas. This included testing prompt variants, identifying tone and style adjustments aligned with FSLs dimensions, and checking for stability in LLM responses. The experimental protocol, task sequence, and data collection procedures were designed manually and finalized in consultation with the research team.
Data Collection	[0% AI] Recruitment of participants, pretesting operations, EEG setup, pupil calibration, and lab coordination. [90% AI] Administration of AI tutoring stimuli during data collection. Generative AI was used to deliver the tutoring responses to participants through system level prompts designed prior to the experiment. [0% AI] No generative AI system made procedural or methodological decisions during data collection.
Statistical Analysis	[20% AI] Generative AI was used to verify sections of statistical code and check the interpretation of model outputs before validation with the lab statistician. All analyses, model selection, preprocessing decisions, and result reporting were conducted by the author and reviewed by the research team.
Thesis writing	[25% AI] Generative AI was used to support several writing related tasks, including grammar correction, paragraph level editing, structural refinement, and improving clarity and flow. It was also used to help translate selected passages from English to French. Drafting, interpretation, and conceptual contributions were completed manually, and all AI assisted text was reviewed, revised, and validated by the author.