PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à École Normale Supérieure-PSL

Optimizing Medical Training Through Learning Analytics and a Large-Scale Experiment

Optimiser la formation médicale grâce à l'analyse de l'apprentissage et à une expérience à grande échelle

Soutenue par

Erva Nihan Kandemir

Le 09 Septembre 2025

École doctorale n°158

ED3C : Cerveau, Comportement, Cognition

Spécialité

Sciences cognitives

Préparée au

LSCP: Laboratoire de sciences cognitives et psycholinguistique

Composition du jury :

Vanda LUENGO Professeure, Université Sorbonne	*Rapporteur*
Franck AMADIEU Professeur, Université de Toulouse	*Rapporteur*
Han VAN DER MAAS Professor, University of Amsterdam	*Examinateur*
Emilie GERBIER Maîtresse de conférences, Université Nice Sophia Antipolis	*Examinatrice*
Bérengère GUILLERY Directrice d'études, École Pratique des Hautes Études–PSL	*Examinatrice*
Franck RAMUS Directeur de recherche CNRS, École Normale Supérieure–PSL	*Directeur de thèse*

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

To my parents,

for their endless love, support, and encouragement,

and to Ömer, with all my love...

Acknowledgements

Reaching the end of this PhD journey feels almost surreal. As I look back, I am filled with deep gratitude for all those who have shaped this path professionally, intellectually, and personally.

First and foremost, I warmly thank my supervisor, Franck Ramus. His constant support, thoughtful guidance, and scientific integrity have shaped every stage of this journey. I am especially grateful for the freedom he gave me to pursue my ideas, the high standards he upheld, and the trust he placed in me. His steady confidence and calm were a constant source of reassurance, and I have learned more than I could have imagined under his guidance.

I sincerely thank the reporters, Vanda Luengo and Franck Amadieu, as well as the jury members, Han van der Maas, Émilie Gerbier, and Bérengère Guillery, for their thoughtful evaluation of this thesis. I am also grateful to my committee members, Émilie Gerbier, François Bauchet, and Jill-Jênn Vie, for following my work over the years and for their comments. I especially thank Jill-Jênn for his curiosity, clarity, and generous collaboration.

I am thankful to the UNESS team, especially Olivier Palombi, Fabrice Jouanot, Adam Sanchez-Ayte, and Maheen Bakhtyar, for their dedication and expertise in bringing the experimental platform to life. This project would not have taken shape without their commitment and collaboration.

Many thanks to the medical students who participated in our studies, and to the CNRS, ENS, and ED3C for their academic and institutional support.

I have always considered myself a lucky PhD student, surrounded by incredible colleagues, supportive lab members, and an office full of green plants and puzzles that made even the hardest days feel a little lighter. For that, I will always be grateful to everyone at the LSCP and especially the Pathology team for creating such a vibrant, stimulating, and welcoming environment that made work both meaningful and enjoyable.

I am particularly thankful to Camille Williams, Emanuele Esposito, Sophie von Stumm (and, of course, our youngest lab member, Clara!), Hugo Peyre, Marie-Christine Lackenbacher, Sophie Bouton, and Ignacio Atal for contributing to such a fruitful and collaborative scientific atmosphere. I also want to warmly thank former members Ghislaine Labouret, who helped me get started with the project, and Alice Latimier, whose advice was invaluable during the meta-analysis phase.

A truly special thank you goes to Lilas, the best labmate I could have asked for, and the greatest gift of this PhD journey. From moments of crisis to everyday routines, you were there with calm, humor, and wisdom. Thank you for your friendship, your intellectual generosity, and your natural ability to make everyone feel included, all of which transformed the lab into a place that felt like home. I will always treasure our lunch-coffee-puzzle rituals, and the way we turned

them into a joyful and grounding tradition.

Thank you to the Cogmaster students, teachers, and interns who joined our team over the past three years. It was a pleasure to work with Alexandre, François, Octave, Elodie, Gabriele, Sophie, Martina, and Iris. A special thank you to Nicolas, the first student I had the chance to supervise, whose curiosity and thoughtful questions taught me as much as I hope I was able to share with him.

I firmly believe that behind every successful woman (if I may call myself one) are exceptional women who walk alongside her. I am deeply thankful to Zeyneb, Seda, and Berfin for being those inspiring women in my life, both in science and beyond. Thank you for your unwavering presence, warmth, and solidarity throughout every stage of this journey.

I could never fully express my gratitude to my parents, Fatma and Mustafa. I am deeply thankful to you for creating such an intellectual home, where I discovered my own way of learning, long before I ever encountered terms like "personalized learning". Perhaps without realizing it, you intuitively respected and supported my individual learning style and laid the foundation for the very topic I explored in this thesis. Even when my progress came in small steps over long periods, you stood by me with unwavering support, and I will always admire you deeply for that.

I also have my siblings to thank, Ayşe and Emin, for always being there and for their steady support throughout these years. I am especially thankful to my sister Neva, who shared this intense final year with me. From late-night pep talks to simply doing the dishes so I could keep writing, your quiet and constant support meant the world.

Finally, my deepest thanks go to my beloved husband and best friend, Ömer, whose enthusiasm continually nourishes my own sense of curiosity. Thank you for your love, patience, and strength, for believing in me even more than I believed in myself, for celebrating every small achievement, and for all the tea breaks filled with scientific conversations. I am especially grateful for the way you were always there to rescue my scripts a million times and to calm my doubts by answering "yes, you did your best" every time I asked, "am I doing okay?". Now comes the big "what is next?" question, and wherever life takes us, I know that with you, the path will be full of learning, purpose, and joy.

Abstract

Digital learning presents a unique opportunity to incorporate evidence-based strategies—such as systematic feedback, retrieval practice, and adaptive difficulty adjustment—while enabling personalized learning through precise monitoring of individual progress. This dissertation investigates the optimization of digital learning within medical education, adopting a cognitive science perspective. Specifically, this research addresses two fundamental questions: the optimization of training difficulty and the effective use of feedback to enhance learning outcomes. To answer these questions, a multi-method approach was employed, combining systematic meta-analysis, learning analytics, and large-scale experimentation on UNESS-BNE, a widely used digital learning system for French medical students.

Chapter 1 introduces the general context and motivation of the dissertation. Chapter 2 focuses on adapting the widely used Elo rating system to estimate both student ability and question difficulty in the UNESS-BNE platform. The results demonstrate that the Elo rating system achieves predictive performance comparable to well-calibrated logistic models in predicting students' final exam outcomes, confirming its suitability for this dataset. Building on this, Chapter 3 utilizes the adapted model within the UNESS-BNE system to examine the optimal level of training difficulty in the context of multiple-choice questions in medical education. The findings support the inverted U-shaped hypothesis, indicating the presence of an optimal difficulty level in this specific learning setting.

Chapter 4 presents a meta-analysis of the effects of feedback timing in digital learning environments. The results indicate that, overall, feedback timing does not significantly influence learning outcomes. However, moderator analyses highlight the impact of factors such as educational level, learning domain, post-test task type, and response time constraints, providing a partial explanation for inconsistencies observed across previous studies. Finally, Chapter 5 details an experimental study investigating the individual and interactive effects of feedback timing (immediate vs. delayed) and initial answer recall on learning outcomes in medical training. Although no significant effects were found, likely due to limited engagement and exposure to the manipulation, data collection is ongoing and may support more conclusive results as it progresses.

Taken together, these studies provide valuable insight into optimizing digital learning in medical education. The findings offer practical implications for practitioners and stakeholders in the broader education sphere, contributing to our understanding of learning, memory, and educational research in digital environments.

Résumé

L'apprentissage numérique offre une opportunité unique d'intégrer des stratégies fondées sur des preuves—telles que le feedback systématique, la pratique du rappel et l'ajustement adaptatif de la difficulté—tout en permettant un apprentissage personnalisé grâce au suivi précis des progrès individuels. Cette thèse examine l'optimisation de l'apprentissage numérique dans l'enseignement médical en adoptant une perspective en sciences cognitives. Plus précisément, elle aborde deux questions fondamentales : l'optimisation de la difficulté des entraînements et l'utilisation efficace du feedback pour améliorer les performances d'apprentissage. Pour y répondre, une approche multi-méthodes a été adoptée, combinant une méta-analyse systématique, l'analyse de l'apprentissage (learning analytics) et une expérimentation à grande échelle sur UNESS-BNE, un système numérique largement utilisé par les étudiants en médecine en France.

Le chapitre 1 introduit le contexte général et la motivation de cette thèse. Le chapitre 2 porte sur l'adaptation du système de notation Elo, couramment utilisé, pour estimer à la fois les compétences des étudiants et la difficulté des questions sur la plateforme UNESS-BNE. Les résultats montrent que le système Elo atteint une performance prédictive comparable à celle des modèles logistiques bien calibrés pour prédire les résultats finaux aux examens, confirmant ainsi sa pertinence pour ce jeu de données. Sur cette base, le chapitre 3 exploite le modèle adapté au sein d'UNESS-BNE afin d'examiner le niveau optimal de difficulté d'entraînement dans les questions à choix multiples en formation médicale. Les résultats soutiennent l'hypothèse en U inversé, indiquant un niveau optimal de difficulté dans ce contexte d'apprentissage.

Le chapitre 4 présente une méta-analyse des effets du moment du feedback dans les environnements numériques. Les résultats indiquent que, globalement, le timing du feedback n'influence pas significativement les performances d'apprentissage. Cependant, les analyses de modération révèlent l'impact de facteurs tels que le niveau éducatif, le domaine d'apprentissage, le type de tâche lors du post-test et les contraintes temporelles de réponse, offrant une explication partielle aux incohérences observées dans les études antérieures. Enfin, le chapitre 5 présente une étude expérimentale portant sur les effets individuels et interactifs du moment du feedback (immédiat vs différé) et du rappel initial de la réponse sur les performances d'apprentissage en formation médicale. Bien qu'aucun effet significatif n'ait été mis en évidence, vraisemblablement en raison d'un engagement limité des participants et d'une exposition insuffisante à la manipulation expérimentale, la collecte de données se poursuit et pourrait permettre d'obtenir des résultats plus concluants à mesure qu'elle progresse.

Ensemble, ces études offrent des perspectives précieuses sur l'optimisation de l'apprentissage numérique dans l'enseignement médical. Les résultats apportent des implications pratiques pour les praticiens et les acteurs de l'éducation, tout en contribuant à une meilleure compréhension des

processus d'apprentissage, de la mémoire et de la recherche éducative dans les environnements numériques.

Publications

Articles included in the dissertation

Kandemir, E. N., Bakhtyar, M., Jouanot, F., Sanchez-Ayte, A., Palombi, O., & Ramus, F. (in prep). *The Effect of Feedback Delay and Initial Answer Recall on Learning.*

Kandemir, E. N., Esposito, E., Gurgand, L., & Ramus, F. (under review). *A Meta-analysis of the Impact of Feedback Timing on Learning Outcomes.*

Kandemir, E. N., Vie, J. J., Sanchez-Ayte, A., Palombi, O., & Ramus, F. (under review). *Investigating the Influence of Training Difficulty on the Learning Outcomes of Medical Students.*

Kandemir, E. N., Vie, J. J., Sanchez-Ayte, A., Palombi, O., & Ramus, F. (2024, March). *Adaptation of the Multi-Concept Multivariate Elo Rating System to Medical Students' Training Data.* In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 123–133). https://doi.org/10.1145/3636555.3636858

Other articles and contributions

Allard, N., Kandemir, E. N., & Ramus, F. (2024). *Scoring systems for multiple-true-false multiple-choice questions: Relationship between student level and actual grade, on an e-learning platform for medical students.* Master's thesis, Cogmaster, ENS.

Contents

Acronyms

ALS Adaptive Learning Systems

BNE Banque Nationale d'Entrainement

EBE Evidence-Based Education

ECN Épreuves Classantes Nationales

EDN Epreuves Dématérialisées Nationales

EdTech Educational Technology

IRT Item Response Theory

LMS Learning Management System

RCT Randomized controlled trials

SIDES Système d'Information pour le Suivi des Étudiants en Santé

UNESS Université Numérique en Santé et Sport

ZPD Zone of Proximal Development

Chapter 1

General Introduction

The objective of this chapter is to provide the foundation for this dissertation by outlining the rationale, theoretical framework, and context of the research. It begins by examining the role of evidence-based learning in educational improvement and the evolving landscape of digital education. Then it presents the specific context of medical education and the UNESS-BNE medical training platform which serves as the primary setting for this research. Finally, it sets the stage for the research by outlining three key points: tracking learning processes, optimizing training difficulty, and optimizing feedback delivery.

Contents

1 Background and Rationale

1.1 The Role of Evidence-Based Learning in Educational Improvement

Despite the growing advances of modern education, a fundamental question remains: How can we ensure that training practices effectively optimize learning outcomes? Leading international organizations emphasize the vital role of research-informed strategies in optimizing learning. The World Bank's 2024 report, *Impact Evaluations for Education Policy*, gathers a decade of impact evaluations to guide policymakers in decision-making (World Bank, 2024). Similarly, the OECD's *Strengthening the Impact of Education Research* supports developing countries in integrating education research into policy and practice. However, translating scientific evidence into practical improvements in education requires a structured and systematic approach. Evidence-Based Education (EBE) has emerged as a key paradigm emphasizing the rigorous application of empirical research to evaluate and refine educational tools, training programs, and learning strategies (Slavin and Cheung, 2017; Slavin, 2020).

A growing number of educational policies and reforms worldwide reflect this commitment to evidence-based approaches. For example, in response to concerns about declining reading proficiency, multiple U.S. states have introduced Science of Reading initiatives (2023–2024), building on earlier federal efforts such as the National Reading Panel report (2000) and No Child Left Behind (2001), which emphasized evidence-based reading instruction. The UK's Education Endowment Foundation (EEF) (2016–present) reinforces the integration of evidence-based teaching practice, while the Netherlands' Education Catch-Up Plan (2021) allocated targeted funding for research-driven interventions to address learning loss following the COVID-19 pandemic. Similarly, France has also implemented large-scale policies informed by empirical research. A notable example is the class size reduction policy introduced by the French Ministry of Education in 2017 to address educational disparities. This policy reduced the size of the first and second grade classes in disadvantaged areas from 24 to 12. Before its implementation, the Ministry conducted evaluations (e.g., Bressoux et al. (2019)) to assess its potential impact on academic achievement within the French context, drawing on accumulated evidence from the broader literature (Filges et al., 2018). In order to further enhance evidence-based policymaking, France established the Conseil Scientifique de l'Éducation Nationale (CSEN) in 2018, marking a major step toward integrating research into education policy. Building on this foundation, France launched Innovation, Data, and Experiments in Education (IDEE) in 2022. This eight-year program (2022–2030) is funded by the Programme d'Investissements d'Avenir and aims to strengthen the use of empirical evidence in education. IDEE supports robust impact evaluations, facilitates access to administrative data, and provides research resources to inform policy and practice.

Cognitive psychology plays a central role in the EBE movement. Since learning is fundamentally a cognitive activity, insights from psychology are essential for understanding and improving key processes such as attention, memory, concept formation, language, metacognition, and problem-solving. As Andler (2008) aptly states, *cognitive science is to education what biology is to medicine*, providing a foundational and methodological framework for EBE. This thesis builds upon this principle, applying empirical research techniques from a cognitive science

perspective to enhance training practices and optimize learning outcomes.

1.1.1 Research Methods and Practical Challenges in Evidence-Based Education

According to Slavin and Cheung (2017), the EBE movement covers four key activities:

1. Identifying the most effective learning practices through well-designed studies.
2. Communicating research findings to educators and policymakers.
3. Providing incentives and resources to support evidence-based teaching methods.
4. Developing policies and systems to expand knowledge on effective educational practices.

The first step in EBE—identifying and evaluating effective learning practices through rigorous research—is the central objective of this thesis. To achieve this, Randomized controlled trials (RCT)s are widely recognized as the gold standard for establishing causal relationships. In educational research, RCTs involve randomly assigning students, teachers, classrooms, or entire schools to experimental programs and comparing them with control groups over extended periods, often spanning a semester or longer, using validated achievement measures.

Despite their strengths in establishing causality and minimizing bias, EBE has been criticized for its heavy reliance on RCTs (Deaton et al., 2018; K. Morrison, 2020). Critics argue that RCTs often fail to account for unmeasured confounders and frequently rely on convenience samples, limiting their generalizability (Deaton et al., 2018; Cartwright, 2020). To address these concerns, researchers advocate for a more flexible approach that incorporates alternative study designs alongside RCTs.

One such alternative is the quasi-experimental approach, often favored by researchers who prioritize statistical control over design control (Cook, 2002; Cook, 2007). Quasi-experimental studies employ techniques such as matching, pretests, and covariate adjustments to minimize pre-existing differences between groups. When these differences are small and relevant confounders are accounted for, quasi-experiments can yield results comparable to those of RCTs.

Yet even the most rigorously designed studies, whether randomized or quasi-experimental, are only as useful as their ability to inform practice across diverse educational contexts. This recognition has led to a broader shift in the EBE movement: from a singular focus on internal validity to a dual emphasis on replication and contextual adaptation. While replication enhances confidence in findings and ensures their robustness, adaptation allows researchers to assess how well evidence-based models perform in different educational environments, uncovering new challenges that may arise. For example, the U.S. Institute for Education Sciences (Institute of Education Sciences, 2018) launched an initiative to fund studies replicating successful educational programs across a broader range of schools and districts.

However, not all studies replicate, and findings may vary depending on context, methods, or sample characteristics. To draw meaningful conclusions from a body of research and make policy recommendations, studies must first be systematically synthesized to ensure that decision-making is guided by the most reliable evidence. Meta-analyses and systematic reviews serve as

key tools in this process. Meta-analysis, by statistically combining results from multiple studies, helps establish the overall strength and consistency of an effect. Systematic reviews, on the other hand, comprehensively gather and assess all relevant studies to generate a broader, generalizable understanding of an educational intervention's effectiveness (Allen et al., 2023; Ahn et al., 2012). Without these synthesis methods, individual studies may be overemphasized or misinterpreted, leading to fragmented or misleading conclusions. Thus, research should include systematic reviews to facilitate informed decision-making in EBE.

Beyond assessment limitations, critics argue that EBE's emphasis on generalizable findings overlooks the complexities of real classrooms. Many evidence-based strategies are developed under controlled conditions, free from the logistical and resource constraints of conventional classrooms, which are inherently heterogeneous, with students varying in prior knowledge, socioeconomic background, and cognitive abilities (Marzano, 2001; Bryk et al., 2015). Consequently, structural limitations in traditional education make the large-scale implementation of research-driven interventions challenging. For example, while France's class-size reduction policy aimed to narrow educational disparities, its implementation required substantial investment in teachers and infrastructure. In practice, however, the policy faced important limitations: the teachers received no specific training, and the measured improvements in student outcomes were modest (DEPP - Direction de l'Évaluation, de la Prospective et de la Performance, 2022). As a result, translating such evidence into policy, despite its potential benefits, may be prohibitively expensive and less feasible at scale. These time and resource constraints ultimately reduce the practical impact of innovative approaches and limit their ability to improve learning outcomes effectively.

Another challenge in evidence-based education is effectively communicating research to educators and policymakers, particularly in the rapidly evolving landscape of education. These stakeholders require reliable, evidence-based tools, yet translating research into actionable strategies remains difficult. Miscommunication or selective interpretation of research findings often leads to the persistence of ineffective methods. A notable example is the 'learning styles' myth, which, despite extensive research showing no benefit—and even potential harm (Clinton-Lisell et al., 2024; Aslaksen et al., 2018; Newton et al., 2020)—continues to influence some educators and policymakers (Furey, 2020). Bridging the gap between research and practice requires not only clear communication but also robust strategies for translating evidence into usable, context-sensitive interventions.

Overall, this thesis adopts a multi-method approach that combines RCTs with complementary methodologies to provide a more comprehensive and nuanced understanding of effective learning and training practices. By integrating adaptation, quasi-experimental studies, literature synthesis, and experimental research, it examines how evidence-based practices can be adapted, evaluated, and optimized to enhance learning in digital medical education environments.

The next section first examines the evolving landscape of learning, driven by technological advancements, and its implications for evidence-based education research before introducing the context of this study—digitalized medical education.

1.2 The Changing Landscape of Learning

Traditional education systems often use passive learning methods and uniform teaching, applying the same instruction to all students regardless of individual differences. The most evident example is the age-based grouping of students based on the assumption that those of similar ages have similar abilities and thus need similar teaching (D. Lee et al., 2022). However, students enter school with widely varying cognitive skills and socioeconomic backgrounds, influenced by a complex interplay of genetic factors and environmental conditions (Bradley et al., 2002). These differences shape both the rate and effectiveness of learning. Despite this variability, traditional education systems often follow fixed curricula and rigid pacing, which fail to accommodate diverse learning needs. As a result, students may experience reduced engagement, weak understanding, and gaps in comprehension and retention (Dumont et al., 2023).

One major problem with this rigid system is the way standardized tests, such as annual exams, are used. While they can provide valuable data for teachers and policymakers, their infrequent and generalized nature makes them less effective for guiding students. When used primarily for feedback or incentives, they prioritize summative over formative assessment, delaying necessary adjustments to learning strategies and limiting opportunities for continuous improvement.

While digital technology does not solve the broader challenge of translating research into practice, it offers promising ways to address some of the structural limitations of traditional teaching. Over the past two decades, advancements in technology-enhanced learning environments have driven rapid growth in Educational Technology (EdTech) (Kew et al., 2022; K. Zhang et al., 2021). This expansion has been particularly evident during and after the COVID-19 pandemic, which accelerated the adoption of digital learning tools worldwide (Koh et al., 2022; Krutka et al., 2024).

According to UNESCO's 2023 report on technology in education (Antoninis et al., 2023), digital platforms have become a central component of education systems, widely used by students, educators, and institutions. The report highlights the dramatic rise in Massive Open Online Courses (MOOCs), with enrollment growing from zero in 2012 to over 220 million in 2021. Similarly, the language-learning platform Duolingo reported 20 million daily active users in 2023, and by 2022, nearly 50% of lower secondary schools worldwide had internet access for educational purposes. Even in early childhood education, digital tools are increasingly prevalent, with 71% of preschool and kindergarten educators incorporating tablets into their classrooms (Dore et al., 2020).

1.2.1 Personalized Adaptive Learning

The concept of personalized learning dates back to early educational philosophies, such as Confucius's principle of teaching students according to their aptitude and Socrates's elicitation teaching method. Historically, personalized instruction was implemented through specialized teaching structures, such as private tutoring in ancient education systems and, more recently, elective course models. However, traditional personalized learning approaches are inherently

constrained by resource limitations, making it difficult to deliver fine-tuned, real-time, and precise adaptations to each learners' needs.

As educational technologies continue to evolve, new solutions have been developed to meet the changing needs of learners. One of the most significant advancement is the integration of Learning Management System (LMS), which are web-based systems that utilize both synchronous and asynchronous technologies to deliver educational content to support instruction and assessment (Turnbull et al., 2019). Examples of widely used LMSs include Moodle, Blackboard, Canvas, and Google Classroom, each offering various tools for content management, communication, and learner assessment. However, while LMSs effectively organize and distribute learning materials, they often lack the ability to tailor instruction to individual learners. To address this limitation, Adaptive Learning Systems (ALS) (J. Lee et al., 2008; Aleven et al., 2016) have emerged as a complementary innovation by leveraging learners' prior interactions and dynamically adjusting content to match individual preferences and learning needs. To address the evolving needs of learners, these educational technologies enable education to move beyond a one-size-fits-all approach, allowing for personalized learning pathways that enhance the applicability of EBE findings across diverse educational settings (Redding, 2013).

Before exploring the role of personalized learning through ALS and Learning analytics (LA) in evidence-based education, it is important to clarify these concepts. Personalized learning encompasses a broad range of instructional approaches and is often used interchangeably with adaptive learning. However, Peng et al. (2019) distinguishes the two: adaptive learning involves dynamically adjusting instruction based on a learner's characteristics and performance, whereas personalized learning considers both cognitive factors and personal development. In this sense, adaptive learning is a key component of personalized education, using learner modeling to tailor instruction to individual needs. Despite this distinction, the terms are often used interchangeably in the literature (Major et al., 2021; Bernacki et al., 2021), highlighting their conceptual overlap. Accordingly, this thesis adopts a flexible use of both *personalized learning* and *adaptive learning* to refer to instructional systems that customize content and pacing based on a learner's abilities, preferences, and progress.

Recent literature strongly supports the effectiveness of personalized learning, particularly when integrated with Adaptive Learning Systems (ALS). Personalized learning has been shown to enhance motivation, foster positive learning attitudes, and develop metacognitive skills and self-reflection. Studies indicate that personalized spacing (Voice et al., 2020), adaptive feedback (Zheng et al., 2022), and personalized recommendations (Karaoglan Yilmaz et al., 2022) lead to greater student achievement than traditional, non-adaptive instruction (Plooy et al., 2024; Tabibian et al., 2019; Mirari, 2022; Deunk et al., 2018). Beyond academic achievement, adaptive learning has demonstrated a significant positive impact on student engagement (Yaseen et al., 2025; Contrino et al., 2024). Higher levels of personalization are consistently associated with improved academic performance, school culture, and student engagement (McClure et al., 2010).

Meta-analyses highlight the potential of ALS to improve learning outcomes compared to non-adaptive methods, particularly in domains suited to individualized practice and feedback. Research on intelligent tutoring systems, a subset of ALS, shows substantial learning gains, with

reported effect sizes ranging from $d = 0.6$ (Z. Xu et al., 2019) to $d = 0.76$ (VanLehn, 2011). However, it is important to contextualize these findings: the control conditions in such studies often include static computer-based instruction, teacher-led large-group instruction, or textbook use—forms of teaching that may not represent the full spectrum of "traditional instruction". For instance, Ma et al. (2014) found that intelligent tutoring systems can outperform these conventional approaches and achieve effects comparable to individualized human tutoring. Yet, such systems primarily excel in well-defined, procedural domains and are best viewed as complements to, rather than replacements for the broader educational experience. This is in line with guidance from the Education Endowment Foundation (EEF, 2021), which cautions that the effectiveness of digital learning depends heavily on thoughtful implementation, teacher involvement, and alignment with curriculum goals.

While much of this research has been conducted in developed countries, a meta-analysis by Major et al. (2021) examined studies from 2007 to 2020 to assess the impact of personalized learning technologies in low and middle-income countries. The findings indicate that adaptive learning positively influences student outcomes, with the most substantial benefits observed in approaches that adapt learning to students' proficiency levels, rather than those solely focusing on interests or feedback.

Similarly, another meta-analysis by (Zheng et al., 2022), covering research from 2011 to 2020, found that adaptive learning had a moderate effect on students' academic performance and a smaller effect on their perceptions of learning. The study emphasized the role of moderating factors, such as learning strategies and software design, in determining the effectiveness of personalized instruction.

More recently, Hu (2024) conducted a meta-analysis on artificial intelligence-assisted personalized learning, finding moderate positive effects on student learning outcomes in terms of knowledge acquisition, competency development, and emotional growth. Additionally, variables such as the type of EdTech applications used, the learning environment, and the duration of implementation were found to significantly influence the impact of AI-driven personalized learning on student success.

With this evidence, there is a clear reason why education is increasingly moving away from standardized, one-size-fits-all instruction toward personalized education with the help of ALS. As EdTech continues to reshape education and learner behaviors are tracked more precisely than ever, it offers new opportunities to apply and extend traditional Evidence-Based Education (EBE) methodologies. In addition to controlled experiments, researchers can now leverage digital platforms to conduct large-scale studies and extract meaningful insights from real-world learner data. Likewise, educators and policymakers can adopt data-informed decision-making to refine instructional strategies based on individual learning profiles (Conole et al., 2011). In this way, technological advancements enrich the application of EBE and help address the complexities of modern learning environments.

The next section explores learning analytics as a key tool in evidence-based education within this fast-changing digital world.

1.2.2 Learning Analytics

In 2011, Society for Learning Analytics Research (SoLAR) defined learning analytics as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" at the first International Learning and Knowledge Conference (LAK) (Conole et al., 2011).

Since then, the field has evolved significantly, enhancing our understanding of learning processes. More recently, Wise (2019) described learning analytics as the application of data science techniques designed to meet the specific demands and challenges of educational settings. Learning analytics now supports the development of ALSs, predictive models, and targeted interventions. Dawson et al. (2014) emphasizes its role in both formative and summative assessments, providing real-time feedback for students and instructors while guiding curriculum design and teaching decisions. As technological advancements continue to evolve, learning analytics ensures that education remains dynamic and responsive to the needs of modern learners, optimizing outcomes in real-world educational settings.

This section highlights three key opportunities that digital learning systems provide through the integration of learning analytics:

1. Enhancing data collection and quality.
2. Serving as a platform for embedded learning experiments.
3. Facilitating the practical application of educational research findings.

1. Enhancing data collection and quality: Digital learning systems are not only tools for content delivery but also powerful mechanisms for collecting, analyzing, and interpreting vast amounts of student-generated data. By tracking student interactions through back-end log data, these systems provide researchers with valuable insights into learning behaviors (Chanifah et al., 2021). This advancement has expanded data collection beyond traditional administrative records to include real-time engagement metrics, such as time spent on tasks, interaction patterns, and participation in discussions (Winne, 2017). Moreover, the application of learning analytics techniques—such as clustering, sequential analysis, and predictive modeling—enables researchers to classify learners based on behavioral patterns. This deeper understanding of student behavioral patterns facilitates more precise analyses of learning processes and supports the development of tailored, data-driven educational interventions (Wintoro et al., 2022).

2. Serving as a platform for embedded large-scale learning experiments: Evidence-based education has increasingly been associated with large-scale experimental studies. However, within cognitive psychology, the predominant approach for evaluating learning interventions is randomized controlled experimentation conducted in controlled laboratory environments. In these studies, learners are placed in isolated cubicles, presented with artificial learning materials, and subsequently assessed on their retention and generalization of the material (Pashler, Bain, et al., 2007; Henry L Roediger III et al., 2012). While these findings are often advocated for

application in educational settings, the disconnect between laboratory conditions and real-world learning environments contributes to a gap between research and educational policy which is the main opposition against evidence-based education. As a solution experimental psychologists argue that validating research findings in authentic educational contexts is essential for ensuring their practical applicability (Boyle, 2012; Koedinger, Booth, et al., 2013).

The practice of conducting learning experiments outside traditional psychology laboratories has a long-standing history, dating back to the 1890s (Bryan et al., 1899; Hall, 1891). This methodological approach is now known as embedded experimentation, which refers to conducting experiments in real educational settings, such as classrooms or informal learning environments by using authentic learning materials and assessment tools that align with pre-existing learning objectives (Motz et al., 2018).

Embedding experiments within established learning environments offers several advantages over traditional laboratory studies. First, learners are less likely to feel self-conscious and more likely to use the learning strategies they naturally employ. In contrast, laboratory settings are often unfamiliar and can put learners at a disadvantage by limiting their sense of authority, control, and comfort. This can interfere with accurately assessing their natural learning behavior.

Second, learning processes and outcomes can differ greatly across cultures, schools, and contexts (Alshumaimeri, 2023; Anyichie et al., 2023). Studying learning in real-world settings helps researchers understand how educational interventions work in practice. This approach provides a clearer picture of how learning happens in different environments.

Third, the idea of a learner gaining knowledge alone in a neutral setting is mostly a myth. Learning is always influenced by the environment and interactions with others. By studying interventions within real-world learning contexts, researchers can assess not only direct effects on individual learners but also the broader impact on the learning community. For example, an intervention that promotes peer discussions can improve understanding for both the participants and their classmates (Corrégé et al., 2021). These indirect benefits are hard to observe in a controlled lab, where students work alone in separate cubicles.

So with all these advantages, embedded experiments encourage researchers to compare interventions in real-world settings. These experiments can be conducted across different classes of the same course (Samsonau, 2018) or multiple schools (Fyfe, 2016; Koedinger and McLaughlin, 2016). However, running well-controlled studies across various classes, schools, populations, and regions, and replicating them remains challenging. It requires significant time, effort, and resources, making large-scale implementation difficult in traditional educational settings (K. Morrison, 2019; Slavin, 2002). As a result, most embedded research in education relies on small-scale studies, where a single intervention is tested in one classroom (Arnold et al., 2017). While these small-scale studies still provide valuable insights, they may not fully capture how interventions work in different learning environments.

Digital learning systems provide a solution by providing a powerful platform for facilitating large-scale embedded experimentation (Ismail et al., 2021). Embedding experiments within digital learning systems offers several advantages over conducting them in traditional classroom settings. First, these digital systems allow for precise manipulation of experimental variables

using learning analytics techniques. For example, in feedback timing experiments, delivering immediate feedback in a classroom setting can be challenging, whereas in a digital learning system, feedback timing can be adjusted by simply modifying a system parameter. This level of control enables researchers to implement experimental manipulations more precisely while preserving the natural learning process.

Additionally, embedding experiments in digital learning systems allows for larger studies that span multiple populations simultaneously, supporting more sensitive comparisons and leading a more robust causal inferences with lower costs. These large-scale studies also provide better insights into demographic factors that may correlate with observed effects, and enhance the generalizability of findings.

Recognizing these advantages, many researchers have already applied learning analytics techniques within digital learning systems to evaluate and implement interventions (e.g. Khiat et al. (2022) and Laeeq et al. (2021)). This practice aligns with a broader trend across digital platforms, where experimentation is routinely embedded to inform design and decision-making. In many contexts, such as language learning apps and large-scale web services, A/B testing is widely used to optimize user experiences and outcomes.

One popular example in the educational domain is ASSISTments, a free, university-supported online learning platform that enables researchers to develop student activities incorporating experimental manipulations while collecting real-time classroom data (Heffernan et al., 2014).

3. Facilitating the practical application of educational research findings: So far, it has been shown how digital learning systems and learning analytics play a crucial role in advancing evidence-based education by improving data collection, enhancing data quality, and facilitating embedded learning experiments. The main goal of evidence-based education is to optimize and enhance student learning by gathering insights into learning processes and designing appropriate interventions. To achieve this, educational research must go beyond analyzing learning mechanisms through data and randomized controlled experiments; it must also bridge the gap between theory and practice by implementing systems, predictive models, and interventions that improve learning outcomes in real-world educational settings.

For instance, the established benefits of personalized learning, as detailed in Section 1.2.1, underscore the necessity of moving beyond traditional, uniform educational approaches. However, the practical implementation of personalized learning at scale requires advanced technological tools.

In this context, digital learning systems and learning analytics have transformed how research findings are applied in practice (Goomas et al., 2021). These modern learning technologies can dynamically adjust the pace of instruction, allowing students to determine when and how they engage with content. Moreover, these systems enable content customization to align with learners' preferences and cultural contexts (Kucirkova et al., 2021) and leverage artificial intelligence to automatically track and respond to students' learning patterns (Boulay et al., 2018).

Given these advantages, ALSs and learning analytics provides a unique opportunity for empirical research, allowing educators and researchers to systematically design, implement, and

evaluate educational interventions. This capability significantly contributes to the advancement of learning sciences. Today, prominent learning platforms such as Duolingo, ALEKS, Coursera, and Khan Academy exemplify the approach of delivering personalized instruction at scale.

Despite their growing popularity and technological sophistication, digital learning systems are not without limitations. One critical concern is the potential overestimation of their ability to replicate or replace the nuanced role of human educators. While adaptive learning systems and learning analytics can personalize content delivery and provide data-driven feedback, digital platforms often lack the human presence necessary to ensure deep engagement and critical thinking (De Freitas et al., 2015). Moreover, O. Simpson (2013) points out that e-learning should more accurately be termed e-teaching, emphasizing that digital environments must be guided by pedagogical principles rather than technological novelty alone. These challenges raise important questions about the feasibility of fully replacing teachers with technology, and instead suggest a more integrated, hybrid approach.

Another significant issue is the high dropout rate in online and digital learning environments. Although these systems offer flexibility and scalability, many learners struggle with sustained engagement and course completion. Systematic reviews have identified isolation, lack of immediate feedback, and limited support systems as key drivers of attrition in online learning (Rahmani et al., 2024; Bawa, 2016). Furthermore, while learning analytics can detect disengagement, interventions are frequently limited to automated alerts or nudges that may lack the nuance to re-engage learners meaningfully. W. Wang et al. (2019) suggested that learner interaction and community-based support are critical to reducing dropout rates, yet are often insufficiently integrated into digital platforms.

This thesis leverages learning analytics and digital learning systems to investigate its research questions by acknowledging both their potential and limitations. The following section introduces digital learning and learning analytics in medical education and details the UNESS-BNE platform, which serves as the primary system for data collection and experimentation in this research.

2 Context: Medical Education and the UNESS-BNE Medical Learning Platform

2.1 Medical Education

Medical education prepares physicians and healthcare professionals with the knowledge, skills, and ethical values necessary for disease prevention and treatment, health promotion, and medical advancement (Buja, 2019). Although specific pathways vary across countries, it typically unfolds over pre-medical education, undergraduate medical training, and postgraduate specialization, (Louw, 2020).

The nature of medical education is inherently interdisciplinary, as it requires the integration of biological, psychological, social, and cultural factors to provide a comprehensive understanding of health and disease. The learning process in medical training is a complex and multi-stage one that relies on both memory-based learning for foundational knowledge and practical learning to apply information in clinical contexts. Furthermore, medical education encompasses a broad range of medical disciplines, requiring students to develop expertise across multiple fields throughout their training.

The field of medicine is dynamic, involving multiple stakeholders such as students, faculty, patients, and healthcare systems, and is influenced by societal needs, technological advancements, and resource availability. Advancements in technology and therapeutics, the emergence of new diseases, and shifting public health priorities continually drive the need for ongoing learning and professional development. Given this constant evolution of medical knowledge, medical education becomes a lifelong process that requires continuous adaptation (Buja, 2019).

2.1.1 Digital Learning and Learning Analytics in Medical Education

Traditionally, medical education has relied mostly on face-to-face teaching methods (Lujan et al., 2006). However, the use of digital platforms in medical education had been expanding for many years, and the COVID-19 pandemic markedly accelerated this trend. Therefore, these digital tools, such as Adaptive Learning Systems (ALS), have also become essential for medical education.

Studies highlight a significant shift toward the use of digital resources among medical students. A survey of 1,626 UK medical students during the COVID-19 lockdown found that 41.6% relied on university resources, 29.6% used free online materials, and 18.4% accessed paid platforms (Barton et al., 2021). Similarly, the 2022 Association of American Medical Colleges (AAMC) survey reported that approximately 70% of students regularly used non-institutional online content for medical education (Medical Colleges, 2022).

These findings highlight the growing preference among medical students for supplementing university materials with digital resources. These digital learning platforms are increasingly favored due to their association with higher exam performance, the ability to self-assess knowledge gaps, improved retention through repeated exposure, and enhanced exam preparation. Therefore, these tools offer significant advantages in efficiency, affordability, and accessibility (Dost

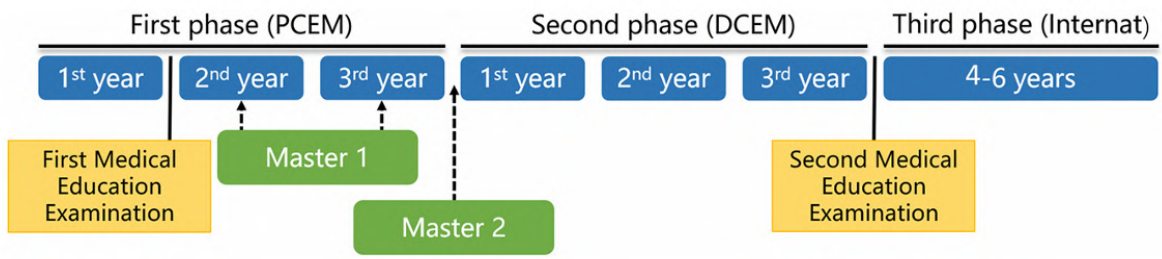


Figure 1.1: Adapted from X. Chen et al. (2024), presents the timeline of medical education and medical science master's training in France. PCEM:premier cycle des études médicale; DCEM: deuxième cycle des études médicales.

et al., 2020). A meta-analysis evaluating the efficacy of online versus offline learning in undergraduate medical education found that online learning offers advantages in enhancing students' knowledge and skills, suggesting its potential as a learning method in medical teaching (Pei et al., 2019). However, while many students found digital learning tools easy to use and beneficial as a supplement, research suggests that digital learning alone cannot fully replace face-to-face instruction. Instead, blended learning—combining digital resources with traditional lectures—still remains the most effective approach for medical education (Popovic et al., 2018; Bock et al., 2021).

The integration of technology into medical education has also paved the way for the use of learning analytics. By incorporating a data-driven approach into medical educational practices, learning analytics enables the adaptation of students to the fast-evolving medical field more effectively. The application of learning analytics in medical education offers several key advantages that include improved decision-making processes and enhanced learning outcomes (Furlan et al., 2022). By enabling personalized and adaptive learning, learning analytics tailors educational experiences to individual student needs and learning rates, ultimately optimizing knowledge acquisition (P. Tanaka et al., 2021). Additionally, learning analytics helps improve evaluation and feedback by providing real-time insights that allow educators to adjust teaching methods and help students track their progress (Cirigliano et al., 2020).

2.1.2 Medical Education in France

France has a long history of medical education focused on training highly skilled medical professionals. The system is known for its long duration and strict selection process, particularly in the first year, where only a small percentage of students progress directly into medical programs due to a highly competitive national examination administered at the end of this initial year (Segouin et al., 2007; X. Chen et al., 2024).

French medical education spans ten to twelve years and is divided into three distinct phases that is also shown in Figure 1.1:

- **First Phase – Premier Cycle des Études Médicales (PCEM) (3 years)** The first year of PCEM serves as a preparatory year for medical studies which covers fundamen-

tal scientific courses. At the end of this year, students must pass the Première Année Commune aux Études de Santé (PACES) exam, a highly competitive national selection examination conducted by the Centre National de Gestion (National Management Center of the French Ministry of Health). Only a limited number of students advance to the second year based on their PACES ranking. Then, the remaining two years of PCEM focus on basic medical sciences in preparation for more specialized clinical training.

- **Second Phase – Deuxième Cycle des Études Médicales (DCEM) (3 years)** During this phase, students transition to a more comprehensive and clinically focused curriculum. They undertake in-depth professional medical studies, gaining practical experience through hospital rotations. At the end of DCEM, students are asked to take the second national medical examination, previously known as Épreuves Classantes Nationales (ECN). This exam determines their ranking, which directly influences their choice of medical specialty and residency placement. However, after nearly two decades of its implementation, the Ministry of Education and Health recognized that the ranking system in ECN often led some students to retake the exam to improve their ranking and secure a preferred specialty or location, particularly in major cities. This leads to an uneven distribution of medical professionals and shortages in rural areas. To address this issue, the ranking system was replaced in 2024 with a matching-based approach, and the exam was renamed Epreuves Dématérialisées Nationales (EDN). The new system includes practical assessments, such as oral exams (Objective Structured Clinical Examination (OSCE)), providing a more comprehensive evaluation of medical candidates (Ministère de l'Enseignement supérieur, 2021b; Ministère de l'Enseignement supérieur, 2021a).
- **Third Phase – Residency Training (Internat) (4 to 6 years)** The final phase of medical education in France consists of residency training, which lasts between four and six years, depending on the chosen specialty. Students who pass the second national medical examination select from 44 nationally defined medical specialties and begin their training at an affiliated university hospital. The number of available residency positions and their locations are determined annually by the French Ministries of Education and Health.

From the second to the sixth year of medical school, including the final national exam, all French medical students take their assessments as multiple-choice questions on tablets connected to a digital platform called UNESS-BNE. This thesis is centered around this digital learning platform, with the aim of optimizing learning outcomes within its framework. Before presenting the research questions and their justifications, the next section will provide a detailed overview of the platform's functionality.

2.2 UNESS-BNE Medical Learning Platform

UNESS-BNE is an online learning platform for medical training in France, managed by Université Numérique en Santé et Sport (UNESS). Originally developed under the name Système d'Information pour le Suivi des Étudiants en Santé (SIDES) and used at the Faculty of Medicine

in Grenoble before being expanded to all 32 medical schools across the country. Since 2014, it has been managed at the national level as a national project.

The SIDES platform was primarily used to administer all medical school exams on tablets. Since 2016, both the ECN and EDN exams have been conducted digitally through this system. It also served as a national training repository, where professors contribute exam content based on a standardized national curriculum organized by medical specialties. This curriculum, established by the French Ministry of Higher Education, is outlined in the Bulletin Officiel (French Ministry for Higher Education and Research, 2013). This shift from paper-based to digital exams has improved efficiency and interactivity in medical assessments by enabling multimedia elements such as images, videos, and sounds while also allowing for automatic grading.

In addition to hosting all official exams, the platform permanently stores past exams, along with numerous additional training questions and clinical cases created by faculty. It provides students with free access to the national training bank- Banque Nationale d'Entrainement (BNE), which contains a vast collection of multiple-choice and single-answer questions designed to prepare them for both national examinations (ECN and EDN) as well as their university exams.

Moreover, the platform fosters student collaboration through interactive features such as chat functions and comment systems. Each day, approximately 6,000 students engage with the platform to supplement their lectures and textbook reading, testing their knowledge using the BNE.

2.2.1 OntoSIDES

All of these interactions with the platform are recorded, and contribute to the creation of the OntoSIDES archive -an ontology-based data warehouse that systematically records all training activities and results of medical students across France since 2013 (Palombi et al., 2019). This data is mapped and linked to elements of the educational curriculum, forming a comprehensive knowledge graph and described at a fine-grain domain ontology level. This extensive OntoSIDES knowledge base stores both past and current student activity, offering integrated access to valuable information for tracking student progress. It is equipped with a powerful query language that allows researchers to explore and analyze specific data, making SIDES a key resource for learning analytics.

In August 2021, the SIDES platform was absorbed by Université Numérique en Santé et Sport (UNESS), a joint initiative (Groupement d'intérêt public) involving 42 universities and the Conférence des Présidents d'Universités. This new version of the platform serves all faculties in medicine, pharmacy, odontology, midwifery, and sports sciences across France. This environment integrates a suite of online applications designed for students, educators, and academic staff in health and sports disciplines.

The platform is organized around three key pedagogical areas:

- **Evaluation**: UNESS Evaluation, UNESS Studio, UNESS Compétences
- **Learning**: UNESS Formation, UNESS Livret, UNESS Portfolio
- **Training**: **UNESS BNE**, UNESS Banque d'Annales (Past Exam Archives)

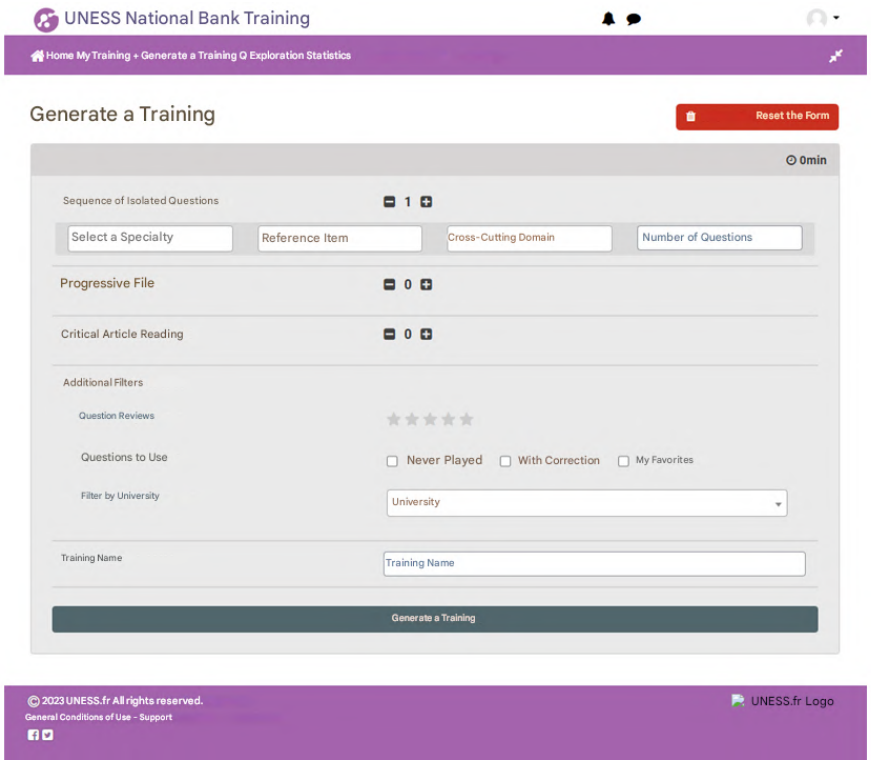


Figure 1.2: English-translated interface of the training generation module on the `uness.fr` platform

2.2.2 Training Module- UNESS BNE

The training module within the platform[1] allows students to engage with the BNE by creating and completing personalized training assessments. The platform offers three distinct types of training exercises that are continuously updated and expanded over time:

- **Isolated Question Sequences (IQS):** Independent, knowledge-based questions that assess specific concepts.
- **Progressive Cases (PC):** Clinical case studies where about 15 pieces of clinical information about a patient and corresponding questions are presented progressively in a fixed order, mimicking real-world diagnostic reasoning.
- **Critical Article Readings (CAR):** Exercises designed to evaluate comprehension and critical thinking skills based on a scientific article that students must analyze.

These exercises are drawn from past national examinations and feature a variety of question formats, including multiple-choice, single-choice, and short open-ended questions. Students can either train using existing exams and training sessions created by professors or other students via

[1] **Terminological note:** In the following chapters, the term **"BNE platform"** refers specifically to the training module of the UNESS system—also referred to as **"UNESS-BNE"**. This simplified terminology is used for clarity and consistency, and should be understood as denoting the training component within the broader UNESS infrastructure.

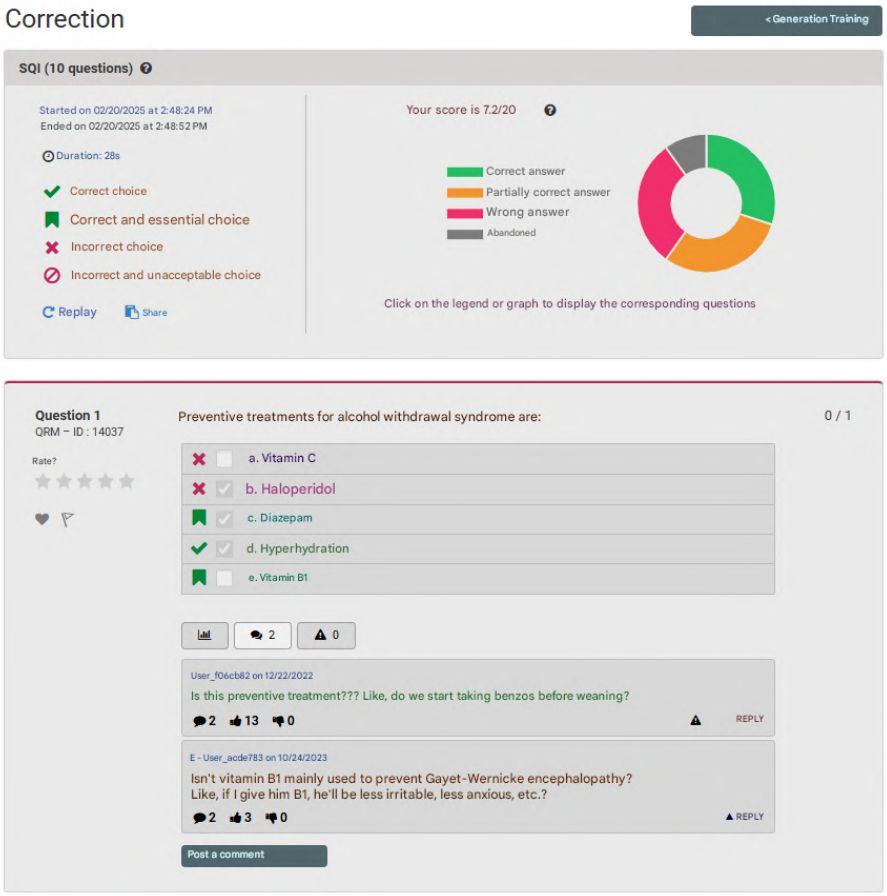


Figure 1.3: Translated version (English) of the correction report generated on the `uness.fr` training platform

the "Exploration" rubric or customize their own training sessions using the "Generate a Training Session" rubric Figure 1.2. In the latter case, they can select one or more medical specialties from the 31 available options. For IQS exercises, they must also specify the number of questions they wish to include. However, note that the training rubric is not adaptive, meaning that questions are drawn at random with varying difficulty levels that do not adjust to the student's individual needs. The training sessions are personalized only in the sense that students can choose the specialty and test type they want to practice, but the selection of questions remains random. Each personalized training session is accessible through the student's interface, allowing for flexible, self-paced practice without time constraints.

Once they complete a session, students receive a comprehensive correction report (Figure 1.3). This report provides detailed feedback on each option in single-choice and multiple-choice questions, classifying them as a *correct choice*, *correct and essential choice*, *incorrect choice*, or *incorrect and unacceptable choice*. Additionally, students can interact with the corrections by adding comments, replying to peers, or liking/disliking others' comments, or scoring each question. These scores serve as valuable metadata and can later be used as a criterion for selecting or filtering questions during training session generation, helping identify high-quality or particularly relevant questions. This feedback loop encourages collaborative learning and

discussion. Students can also review their past training sessions and feedback through the "My Training Sessions" rubric and repeat sessions as often as needed to reinforce their learning.

2.2.3 Experimental Module - BNE Expérimentale

As part of this research project, a dedicated experimental module (*BNE-expérimentale*) was integrated into the UNESS platform. Specifically developed to support the experimental framework of this dissertation, the module was designed and implemented over a two-year period in close collaboration with the platform's technical team. Throughout this process, I was actively involved in all stages of development, contributing to the definition of functional requirements and ensuring that the module aligned with the methodological and research objectives of the project. By extending the standard functionalities of the BNE module, this experimental module enables controlled modifications to learner interfaces and experiences, establishing the infrastructure required for conducting controlled, embedded educational experiments within authentic learning environments.

This BNE Expérimentale module runs alongside the standard BNE module, keeping all core features while adding specific enhancements for research purposes. For this thesis research specifically, ethical approval was obtained, ensuring compliance with data protection regulations and ethical guidelines for human research. Participation is completely voluntary, allowing students to choose between the standard BNE version and the experimental module, with the option to switch at any time. Informed consent ensures that participants understand the purpose of the experiments and agree to the use of their anonymized data.

Together with the OntoSIDES data archive, the experimental module makes the UNESS-BNE platform a valuable tool for learning analytics and large-scale educational research. This thesis leverages the capabilities of the UNESS-BNE platform, including BNE-experimental, to explore and optimize the learning experience of medical students.

3 Key Points in Optimizing Learning

3.1 Measuring and Tracking Learning Processes

Section 1.2 reviewed the literature emphasizing the advantages of adaptive learning compared to traditional, one-size-fits-all methods. It also examined how digital learning systems enable the implementation of these adaptive learning approaches.

For adaptive methods to be effective, learning systems must continuously track users' progress—a process known as knowledge tracing. This involves building learner models based on students' performance and interactions with the system. These models are crucial for personalization in learning management systems, as they help determine a learner's proficiency, identify their needs, and ensure that learning materials (such as questions or problems) are appropriately matched to their skill level (Pelánek et al., 2017; Eglington et al., 2023; Martin et al., 2020).

Adaptivity goes beyond selecting suitable learning materials; it also includes features like adaptive feedback and spaced repetition. In the context of adaptive item selection, learner models play a crucial role in tracking a student's skill level (Klinkenberg et al., 2011). However, it is important to note that these models can be customized to capture a broader range of learner characteristics, extending beyond just the measurement of abilities (Chrysafiadi et al., 2013).

3.1.1 Learner Models

Over the years, various learner models have been developed (Pelánek, 2017). The most prominent models can be categorized into three main groups: probabilistic models, logistic models, and rating models.

Probabilistic Models: This approach primarily relies on probabilistic inference using a two-state Hidden Markov Model (HMM), where one state represents a learner's mastery of a concept and the other indicates non-mastery. A key model in this category is Bayesian Knowledge Tracing (BKT), first introduced by Corbett et al. (1994). While BKT has significantly influenced educational data mining, it has limitations when applied to questions that require multiple knowledge components which are the fundamental units of knowledge necessary for a correct response.

Furthermore, a recent study comparing various performance modeling algorithms on real-world datasets (Gersten et al., 1982) found that when applied to large data, both BKT and its extension, BKT+ (Khajah et al., 2016), demonstrate slower training speeds and lower predictive accuracy compared to modern approaches based on logistic regression and deep learning. These findings highlight the need for further advances to improve the efficiency and precision of knowledge tracing models, a challenge that several research groups are actively addressing (Badrinath et al., 2021; Käser et al., 2017).

Logistic Models: This category of knowledge tracing methods is based on logistic functions to estimate and model factors influencing students' knowledge states. A key example is logistic regression models, which transform a student's interaction history into a feature vector to predict

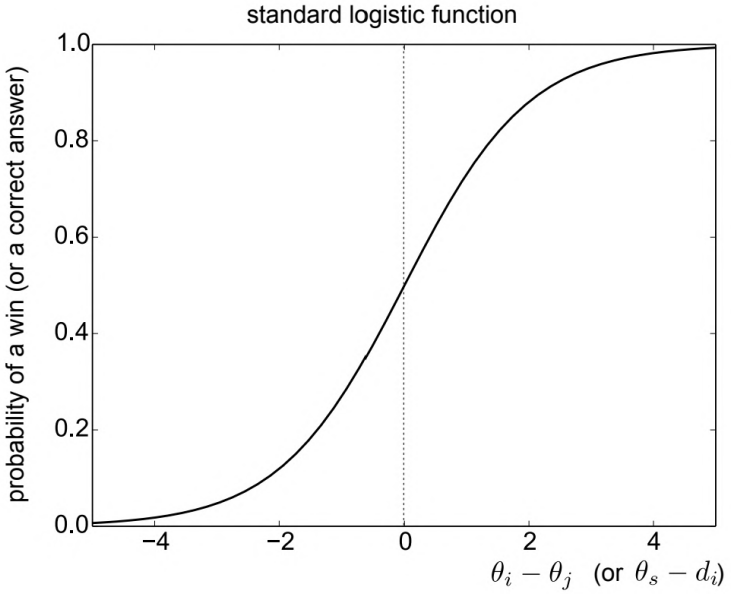


Figure 1.4: The logistic function transforms the difference between opponents' ratings (student skill and item difficulty) into a probability of a win (a correct answer).

the probability of mastering a specific concept or knowledge component. This probability often reflects the likelihood of correctly answering a given question (Figure 1.4). Several techniques fall under this category, including Item Response Theory (IRT) (Linden et al., 2013), Performance Factors Analysis (PFA) (Pavlik Jr et al., 2009), DASH (Lindsey et al., 2014; González-Brenes et al., 2014; Mozer et al., 2016), and its extended version DAS3H (Choffin et al., 2019).

Among these approaches, the IRT framework, specifically the one-parameter version known as the Rasch model (Rasch, 1960), stands out as a foundational simple regression model. It operates by introducing parameters θ_i to denote the ability of student i, and d_j for the difficulty level of question j. IRT's predictive ability lies in its estimation of the probability of a student i successfully recalling question j, denoted as R_{ji}. This estimation is elegantly expressed through a logistic function of the student-specific ability θ_i and the question-specific difficulty d_j:

$$P(R_{ji} = 1|\theta_i, d_j) = \frac{1}{1 + e^{-(\theta_i - d_j)}}$$

To ascertain the model's parameters from observed data, a widely accepted method is the utilization of joint maximum likelihood estimation which is a procedural iterative approach (De Ayala, 2013). It is essential to emphasize that the Rasch model does not account for learning that occurs across multiple attempts and over time due to its absence of temporal input in the prediction formula.

In contrast, PFA (Pavlik Jr et al., 2009) represents an extension of the Rasch model designed to capture the evolution of students' ability over time by taking into account the successive interactions of students with questions. In PFA, a student's ability is expressed as a probability of a correct answer, determined through a linear combination of factors including the item's difficulty and the student's past successes and failures.

Finally, DASH (Lindsey et al., 2014) builds on logistic regression models by incorporating a forgetting curve to improve learning predictions. It dynamically adjusts the estimated probability of recall based on how recently and frequently a student has encountered a concept, modeling how memory decays over time. DAS3H (Choffin et al., 2019) extends DASH by introducing temporal effects and student-specific forgetting rates, enabling more personalized tracking of knowledge retention.

Rating Systems: Rating Systems, particularly the Elo rating system (Elo et al., 1978), have gained widespread use in educational technologies, as demonstrated in various studies (Attali, 2014; Pelánek et al., 2017; Vermeiren et al., 2025; Abdi, Khosravi, Sadiq, and Gasevic, 2019). Originally developed for ranking chess players based on their performance, the Elo rating system has been adapted for educational contexts, where it models both users and learning materials as opponents. In this framework, the Elo rating system assigns ratings to users and items (questions), serving as an indicator of user ability and item difficulty.

For each user u, the parameter θ_u represents their overall ability across the domain, while for each item i, the parameter θ_i represents the item's difficulty. The probability that user u correctly answers item i, denoted as $P(a_{ui} = 1|\theta_u, \theta_i)$, is determined by a logistic function based on the difference between user ability and item difficulty:

$$P(a_{ui} = 1|\theta_u, \theta_i) = \sigma(\theta_u - \theta_i) = \frac{1}{1 + e^{-(\theta_u - \theta_i)}}$$

For the prediction of a correct response, the Elo rating system uses the same logistic function as the Rasch model in IRT, where the probability of success is based on the difference between the user's ability and the item's difficulty.

In the Elo rating system, both user ability and item difficulty are updated after each interaction based on the outcome of the attempt. The updates follow these formulas:

$$\theta_i := \theta_i + K\left(P(a_{ui} = 1|\theta_u, \theta_i) - a_{ui}\right)$$

$$\theta_u := \theta_u + K\left(a_{ui} - P(a_{ui} = 1|\theta_u, \theta_i)\right)$$

where a_{ui} represents the actual response outcome, and K is a constant that controls the sensitivity of updates.

This iterative process ensures that the Elo rating system continuously refines user and item parameters after each interaction. Updates are proportional to the difference between the estimated probability of a correct response and the actual outcome, allowing for a dynamic and adaptive assessment of user ability and item difficulty.

Although various versions of learner models exist, implementing them in digital learning environments presents significant challenges, particularly in terms of computational efficiency and accuracy. Because fitting an IRT model uses the entirety of the data on learners and questions, it requires substantial computational resources for real-time calibration, making it impractical for ongoing adaptation in online learning settings. As the number of users and tasks increases,

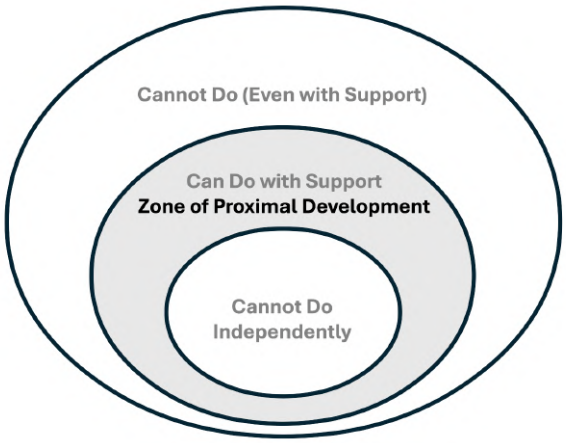


Figure 1.5: A schematic representation of Zone of Proximal Development (ZPD) by Vygotsky.

obtaining rapid estimates becomes increasingly difficult, limiting the feasibility of IRT in dynamic educational systems. To overcome this, alternative methods that enable faster difficulty estimation have been explored, such as the Elo rating system. Although it is computationally more efficient than IRT and designed for real-time updates, its application to different learning environments still requires modifications to account for the unique characteristics of the data and assessment structure (Vermeiren et al., 2025; Abdi, Khosravi, and Sadiq, 2021; Abdi, Khosravi, Sadiq, and Gasevic, 2019).

In this thesis, the Elo rating system was first adapted to the specific dataset of BNE, addressing its unique challenges. After validating its predictive accuracy, the model was then used to explore the broader question of optimal training difficulty, which is introduced in the following section.

3.2 Identifying and Applying the Optimal Level of Difficulty in Learning

Difficulty is an inherent and essential aspect of the learning process, but it must be appropriately calibrated. When engaging in learning—whether acquiring new knowledge and skills or refining existing ones—learners naturally seek this appropriate level of difficulty.

From a motivational perspective, excessively difficult tasks have been shown to lower goal-setting (Horvath et al., 2006), reduce productivity (Goemaere et al., 2018), and relate negatively to learners' perceived competence (Patall et al., 2018). Conversely, tasks that are too easy can lead to boredom, disengagement, and reduced motivation, as learners may not feel sufficiently challenged to stay cognitively involved (Power, 2019). In contrast, studies have demonstrated that selecting an appropriate difficulty level for practice exercises positively impacts learning gains (Sampayo-Vargas et al., 2013), enhances the learning experience (Kostons et al., 2010), increases engagement (Papoušek, Stanislav, et al., 2016b), and enjoyment (Abuhamdeh and Csikszentmihalyi, 2012). Thus, effective learning requires an optimal balance of challenge and feasibility.

The concept of a "sweet spot" in learning where challenge is optimally balanced to enhance motivation and progress, is central to both learning theory and instructional design. This idea

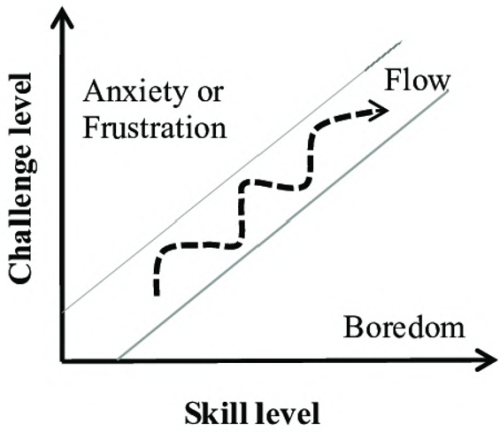


Figure 1.6: A schematic representation of Flow Theory by Csikszentmihalyi (1990).

has been systematically explored and supported by various educational frameworks:

Zone of Proximal Development (ZPD): The idea that learning should be optimally challenging is often attributed to L. S. Vygotsky et al. (1978)'s theory of the Zone of Proximal Development (ZPD). According to this theory, the most effective learning occurs when tasks are slightly beyond the learner's current level of ability that they cannot complete independently but are still achievable with appropriate guidance or support (Figure 1.5). This framework emphasizes the need to balance challenge and achievability so that learners advance without feeling overwhelmed.

Importantly, the ZPD is not a fixed level of difficulty but a range that changes as the learner gains skills. Learning is maximized when tasks consistently fall within this zone, allowing for gradual skill enhancement through guided practice. For example, a child who recognizes letters and simple words like "cat" or "dog" on their own may need support to read short sentences with phonetic words. As their reading ability grows, their ZPD expands to include more complex sentences with multisyllabic words, and later, full paragraphs with abstract vocabulary and advanced grammar. This way, adjusting the level of challenge over time ensures continuous progress and skill development.

Flow Theory: Csikszentmihalyi (1990) introduced Flow Theory, which suggests that optimal learning happens when individuals enter a state of flow (Figure 1.6). This psychological state involves complete absorption in an activity, where actions feel effortless, self-consciousness fades, and time seems to pass unnoticed. A classic example is a musician deeply engaged in playing a challenging piece, experiencing mastery without conscious effort.

In the context of learning, students who are highly motivated, whether intrinsically or extrinsically, may experience this state of flow, where the perceived effort of learning is minimized due to deep engagement. Flow Theory states that this state occurs when the challenge of a task slightly exceeds a learner's ability, creating a sense of progress and engagement. Learners in flow focus intensely and sustain effort while finding the task enjoyable and rewarding. In this state

of flow, they stay focused on their goals and persist through challenges since they are motivated by the intrinsic satisfaction of progress and mastery (Swann et al., 2016). In contrast, if a task is too difficult, learners may experience anxiety and frustration; if it is too easy, they may become bored and disengaged, ultimately losing the flow state.

Fundamentally, Flow Theory and the ZPD share the same core principle: when learners operate within their ZPD, they are more likely to experience a state of flow.

Desirable Difficulties Framework (DDF): While excessive difficulties are usually undesirable because they hinder learning, some degree of difficulty enhances learning by triggering encoding and retrieval processes that are essential for learning. The Desirable Difficulties Framework (DDF), proposed by R. A. Bjork (1994) and R. A. Bjork and E. L. Bjork (2020), suggests that introducing a manageable level of difficulty during learning balances challenge and feasibility and enhances long-term retention. According to this theory, learning tasks should be designed to be sufficiently challenging to promote deeper cognitive processing while remaining achievable. Although desirable difficulties may initially lead to poorer performance during the acquisition phase, this increased effort ultimately enhances long-term retention and facilitates knowledge transfer. However, if learners lack the necessary background, these difficulties become undesirable, as they interfere with rather than support learning. Therefore, the effectiveness of desirable difficulties depends on the learner's prior knowledge, with the appropriate level of challenge varying accordingly.

Cognitive Load Theory (CLT): Cognitive load theory provides another lens through which to understand the optimal level of difficulty in learning by examining the cognitive demands placed on learners during learning. CLT emphasizes the importance of effectively managing those demands to improve learning outcomes (Sweller, 1988). According to CLT, cognitive load can be categorized into three types: intrinsic load, which is inherent to the complexity of the information being learned; extraneous load, which is imposed by the way the information is presented; and germane load, which is related to the mental effort required to process and understand the information.

For optimal learning, CLT suggests that extraneous load should be minimized so that learners can allocate more cognitive resources to intrinsic load (understanding the content itself) and germane load (actively constructing meaningful knowledge structures). This means that learning materials should be designed with an appropriate difficulty to effectively manage the cognitive load, and reduce the risk of experiencing cognitive overload, which would result in frustration.

Inverted U-Shape Hypothesis: Abuhamdeh and Csikszentmihalyi (2012) examined the relationship between difficulty and enjoyment and found that it follows a curvilinear pattern. As difficulty increases, enjoyment also rises until it reaches an optimal point (the peak of the curve). Beyond this point, further increases in difficulty lead to a decline in enjoyment, a pattern described as the inverted U-shape hypothesis (Figure 1.7). This finding supports the idea that the most engaging and optimally challenging activities strike a balance between difficulty and enjoyment.

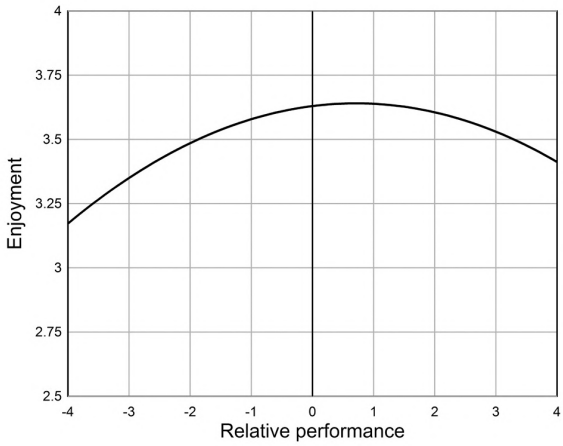


Figure 1.7: Enjoyment as a function of relative performance adapted from Abuhamdeh and Csikszentmihalyi (2012)

Collectively, these theories converge on a fundamental principle: learning is most effective when tasks are optimally challenging—neither too easy nor excessively difficult. However, in practice, traditional teaching typically offers a single level of difficulty, meaning that some students may find tasks too challenging, while others may find them too simple. Both situations affect learning and engagement negatively. Therefore, identifying and maintaining the optimal level of difficulty remains a significant challenge in education.

The first challenge in optimizing learning difficulty lies in accurately assessing the zone of proximal development. Understanding theories that emphasize the balance between engagement and difficulty leads to a fundamental question: what defines this optimal balance between tasks that are too easy and those that are too difficult, and what factors influence it?

Although educators do their best to offer exercises that fall within each student's zone of proximal development, variations in prior knowledge, skills, and learning rate mean that this "zone" differs for every learner. As a result, in large classrooms, identifying and maintaining this optimal difficulty level is particularly challenging due to the absence of a rapid system for tracking students' individual learning patterns. To address this, researchers have explored various strategies to determine the ideal difficulty level.

Since optimal difficulty depends on a learner's prior knowledge, researchers have shifted their focus from absolute difficulty to relative difficulty, which measures how challenging a task is in relation to the learner's ability. As a result, various approaches aim to determine the optimal accuracy rate for a task. These include theoretical models, data-driven clustering techniques (Yuhana et al., 2024), and adaptive learning methods (Gallego et al., 2018).

One prominent finding in this area comes from R. C. Wilson et al. (2019), who studied learning algorithms—particularly those based on stochastic gradient descent—and found that the optimal difficulty level occurs when accuracy is around 85% in a binary discrimination task. This "85% Rule" suggests that learning is most efficient when the error rate is approximately 15%, a conclusion supported by multiple model simulations across binary decision-making tasks.

Similarly, research on effective instructional methods in elementary mathematics classrooms

(Rosenshine, 2012) found that the most successful teaching approaches maintained student accuracy rates in the low-to-mid 80% range (approximately 82–85%). Additionally, in mastery-based instructional models such as Keller's Personalized System of Instruction (Eyre, 2007), mastery is often defined by an accuracy threshold of 85–90% to ensure sufficient retention of prior material before progressing to new content.

These findings align with theoretical perspectives such as desirable difficulty, flow theory, and Vygotsky's ZPD, all of which emphasize that learning is most effective when it presents a challenge just beyond the learner's current ability level. The 85% success rate identified in these studies reflects this principle, suggesting that optimal learning occurs when tasks are demanding enough to promote growth while remaining achievable with effort and active engagement.

However, the optimal difficulty level is not universal but varies according to multiple interacting factors, including the subject matter, the structure and type of task, and the learner's prior knowledge. While studies report similar optimal success rates near 85% as a potential sweet spot for maximizing learning efficiency, this numerical threshold should not be interpreted as representing the same level of challenge across all contexts. For example, a success rate of 85% in a binary multiple-choice task where learners choose between two options involves substantially lower cognitive demands than achieving the same accuracy in open-ended problem-solving, generative writing tasks, or complex procedural learning. Similarly, in the domain of skill acquisition, Al-Fawakhiri et al. (2023) provide both theoretical and empirical support for an optimal success rate of approximately 70% in motor learning tasks.

Moreover, a study by (Papoušek, Stanislav, et al., 2016b) on learning geographical facts found that more challenging questions enhance long-term learning and engagement, while easier questions are more effective for short-term engagement. Their findings suggest a dynamic approach, where learning begins with easier questions to capture learners' interest before transitioning to more difficult ones, rather than relying on a fixed difficulty threshold.

These variations highlight that the ideal difficulty level is highly context-dependent. Therefore, effective learning systems should adaptively set difficulty thresholds based on the subject matter, task type, and learner needs, rather than applying a one-size-fits-all threshold.

The second key challenge in maintaining optimal difficulty level in learning is the fact that during the learning process, learners' knowledge and abilities are not static but evolve over time (Gallego et al., 2018). As a result, the perceived difficulty of learning materials fluctuates. To sustain an appropriate level of challenge, a dynamic approach is necessary—one that continuously tracks learners' evolving knowledge states.

Moreover, achieving optimal relative difficulty requires assessing not only the evolving abilities of students but also the difficulty of each item or question. Since a learner's perception of an item's difficulty is inherently relative to their ability, many studies have focused on efficiently and automatically measuring both question difficulty (Yaneva et al., 2019; Huang et al., 2017) and student ability. In this regard, numerous studies have explored efficient and automated methods for measuring question difficulty and student ability by enabling real-time matching and adjustments.

Within this framework of ability-adjusted item selection, learner models play a central role

by providing real-time estimates of learner ability and item difficulty (Klinkenberg et al., 2011). These estimates are then used to compute the probability of success on potential tasks, guiding the system to select questions that match the learner's current skill level. In computer-based education, this automatic adjustment of difficulty levels has been referred to by various terms, including computer-adaptive practice (Pelánek et al., 2017; Klinkenberg et al., 2011), adaptive curriculum (Belfer et al., 2022), personalized task difficulty (Y. Zhang et al., 2021), and dynamic difficulty adjustment (DDA) (Hunicke, 2005). Despite different terminologies, these methods share a common goal: assessing the complexity of learning tasks in real-time and adjusting difficulty levels to match the learner's needs. Moreover, when implementing these dynamic difficulty adjustment techniques, Adaptive Learning Systems (ALS)s are instrumental, as empirical studies have consistently demonstrated the effectiveness of adaptive item selection in maintaining an optimal level of challenge (Broeke et al., 2022; S. Wang et al., 2023; Segal et al., 2018).

In conclusion, maximizing learning efficiency requires identifying and applying the optimal difficulty level. This involves two essential steps: (1) determining the ideal balance between challenge and engagement to establish a target difficulty level, and (2) continuously estimating both the learner's knowledge and the difficulty of the learning materials to ensure that each selected item meets this predetermined level.

In this thesis, learning models—particularly rating systems—were employed to track learners' knowledge states and assess the difficulty of individual items. Additionally, statistical modeling was used to assess the optimal difficulty level across various medical specialties within the BNE platform.

3.3 Identifying the Optimal Ways to Use Feedback

3.3.1 The role of errors in learning

In the educational context, errors are defined as responses to learning tasks that deviate from either a learner's subjective expectations or an established objective criterion (A. Simpson et al., 2020). They are an inherent part of the learning process that might occur both during initial acquisition and throughout practice and refinement. This raises a critical question: Should learners actively engage with their errors by committing, exploring, analyzing, and correcting them, or should errors be minimized and avoided at all stages of learning?

The traditional approach to learning emphasizes avoiding or minimizing student errors. This perspective is based on early learning and memory theories, which argue that since the goal of learning is to achieve accurate performance in assessments, errors during practice might be reinforced, making them more difficult to correct (Terrace, 1963; Bandura et al., 1986; Skinner, 1965; Ausubel et al., 1978). As a result, this approach promotes "errorless learning" environments, where only correct responses are encouraged, and mistakes are deliberately minimized or ignored.

However, this errorless learning approach has two significant limitations. First, it is impractical, as errors are an unavoidable part of most learning contexts. This is particularly evident

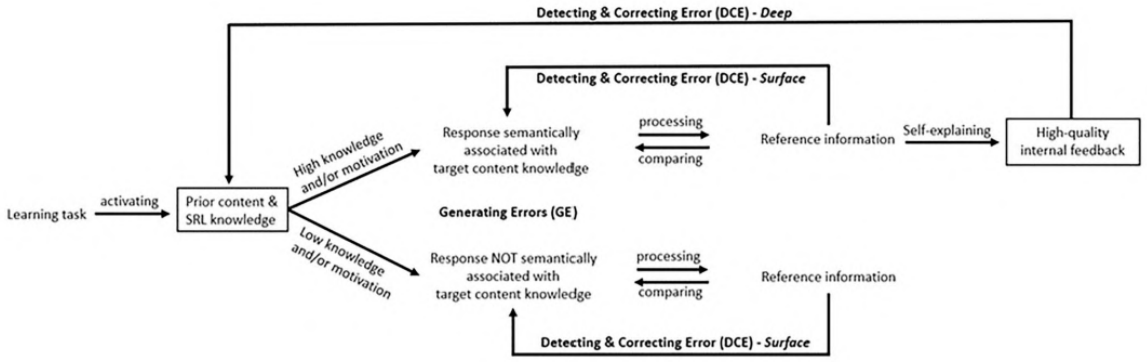


Figure 1.8: Adapted from Qian Zhang et al. (2023), a theoretical model of learning from self-generated errors. SRL: self-regulated learning.

when considering the previous section, which highlights that optimal learning occurs when tasks slightly exceed the learner's current ability. At this level of challenge, errors naturally emerge as a byproduct of engaging with complex or unfamiliar material, making them an integral part of the learning process.

Second, the theoretical foundation of errorless learning is suboptimal, as research has shown that errors can serve as valuable learning opportunities when the appropriate conditions are met. This perspective has long been recognized in the literature. For example, (Piaget, 1952)'s theory of cognitive development emphasized that learning is driven by a state of disequilibrium, in which children encounter obstacles or contradictions—such as errors—due to gaps or inaccuracies in their knowledge. To resolve this disequilibrium, learners must either integrate new information into their existing knowledge structures or modify those structures to accommodate conflicting information. In educational settings, this process is often facilitated through feedback or teacher guidance.

In addition to old learning theories, more recent perspectives also suggest that making and correcting errors can be a powerful learning strategy. Numerous empirical studies have demonstrated that, under specific feedback conditions, generating errors can be more beneficial for learning than errorless learning (Butterfield et al., 2006; Potts et al., 2014; Metcalfe, J. Xu, et al., 2025; Metcalfe and Eich, 2019; Kornell, Hays, et al., 2009); see Metcalfe (2017), Qian Zhang et al. (2023), and Mera et al. (2022) for a review. One striking example is the "derring effect" (Wong et al., 2022), which shows that even when learners already know the correct answer, deliberately committing and then correcting an error can enhance retention.

Studies in educational psychology (Loibl et al., 2018; Loibl et al., 2019; Kapur, 2014; Kapur, 2016) have also demonstrated that engaging with errors under the right conditions facilitates deeper comprehension and improves the ability to apply and transfer knowledge across different contexts.

To explain how errors can facilitate rather than hinder both lower-order retrieval-based learning and higher-order learning, Qian Zhang et al. (2023) proposed a two-stage learning model (Figure 1.8). The first stage involves engaging with tasks and generating errors, while the second focuses on detecting and correcting those errors. In the initial stage, prior knowledge and

self-regulated learning play a central role in task processing and error generation. Depending on a student's level of prior knowledge, the errors they produce may be either semantically related or unrelated to the content. The second stage is where external support, such as feedback or instruction, becomes critical for identifying and correcting errors. When errors are semantically related to the content, students can effectively refine their understanding by comparing their responses with reference information—most commonly provided through feedback.

As a result, given the inevitability of errors in learning, their impact depends largely on how they are treated. In this context, feedback is essential for learning from errors, as students can only correct their mistakes once they become aware of them. Effective feedback plays a crucial role in this process by helping learners identify, understand, and correct errors in ways that improve retention and knowledge transfer (R. C. Anderson et al., 1972; A. C. Butler and Roediger, 2008; Kornell and Metcalfe, 2014). Without feedback, or without the right kind of feedback (Qian Zhang et al., 2023), errors are more likely to persist. The next sections explore the critical role of feedback in learning , explores its underlying mechanisms, and discusses the key factors that influence its effectiveness.

3.3.2 Feedback on learning

Feedback is commonly defined as a mechanism that confirms a learner's performance, provides corrections, or offers suggestions for improvement. Many researchers adopt this general definition with different wording and focus on specificity (e.g., D. L. Butler et al. (1995), Hattie and Timperley (2007), and Narciss (2013)). By definition, feedback is widely recognized as a powerful tool for learning and achievement, as it helps bridge the gap between a learner's current understanding and their desired level of proficiency (Wisniewski et al., 2020; Hattie and Timperley, 2007). Beyond its corrective function, feedback can also support learning in indirect ways. For instance, studies have shown that it can influence learners' emotional states (Mory, 2013; Shute, 2008) and modify their behaviors (Hattie and Timperley, 2007). Additionally, feedback has been linked to changes in motivation (e.g., Koenka et al. (2021) and Fong et al. (2019)).

With the advancement of digital learning systems, the nature of feedback has evolved. Compared to traditional learning settings, these environments offer several advantages, such as enhancing the learning experience (Van Ginkel et al., 2019), providing strategic guidance (Narciss, 2013), and facilitating the construction of meaningful knowledge (Fu et al., 2018). In addition, these systems allow for more personalized, detailed, and immediate feedback (Cai et al., 2023; Brummer et al., 2024). As a result, researchers have increasingly focused on understanding the mechanisms of feedback in these digital environments.

C. F. Timmers et al. (2015) proposed a five-stage circular learning process model of computer-based formative assessment, which describes how feedback influences individual learning in digital learning systems (Figure 1.9). This model highlights the iterative nature of learning and the crucial role of feedback in reducing uncertainty. By decreasing uncertainty, feedback enables learners to adjust their strategies and enhance their subsequent learning. Specifically, it allows learners to refine their approach in real-time, fostering more effective learning outcomes.

Several meta-analyses and review studies have examined the impact of feedback in computer-

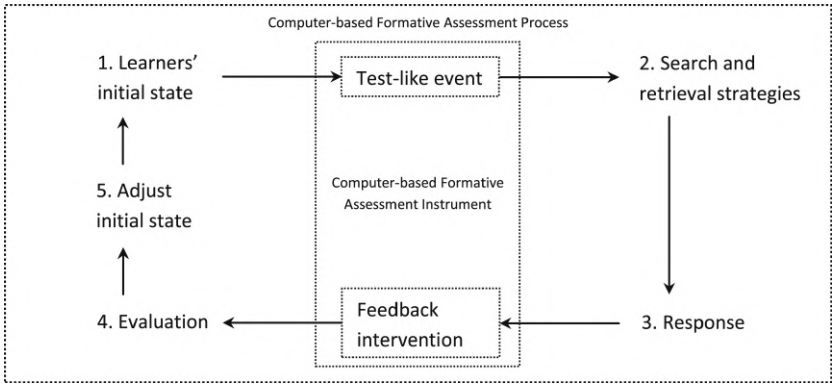


Figure 1.9: Conceptualisation of computer-based formative assessment based on the five-stage model adapted from C. F. Timmers et al. (2015)

based learning environments (Azevedo et al., 1995; Eggen et al., 2011; Wisniewski et al., 2020; Cai et al., 2023; Brummer et al., 2024; Mertens et al., 2022). These studies generally suggest that feedback can significantly improve learning performance in computer-based learning environments. However, its effects can be positive, neutral, or even negative, depending on the nature of the feedback and how it is delivered.

3.3.3 Factors Influencing Feedback Effectiveness

Feedback delivery strategies can be designed in various ways and can differ based on their function (e.g., cognitive, metacognitive, and motivational), content (e.g., evaluative and elaborative components), and feedback characteristics (e.g., modality, frequency, type, channel, timing, and valence of feedback) (Narciss, 2008; Narciss, 2017; Shute, 2008; Hattie and Timperley, 2007). Therefore, it is crucial to explore the optimal conditions under which feedback is most effective in correcting errors and supporting learning. Here, the feedback conditions examined in the context of this thesis are introduced.

Feedback Timing: The timing of feedback is a key question in educational research, particularly regarding whether immediate or delayed feedback leads to better learning outcomes. A seminal meta-analysis by Kulik et al. (1988) found that the effectiveness of feedback timing varied depending on the learning environment. Specifically, laboratory studies tended to support delayed feedback, whereas classroom studies favored immediate feedback.

With the increasing use of digital learning technologies, it is essential to reassess the role of feedback timing, as modern platforms introduce new modalities for delivering feedback that may influence its effectiveness. Several meta-analyses have examined feedback timing in digital learning environments, either by considering it as a moderator variable (Azevedo et al., 1995; Swart et al., 2019) or by categorizing studies post hoc based on whether they delivered immediate or delayed feedback (Van der Kleij, Feskens, et al., 2015). However, there is a lack of systematic reviews that focus exclusively on studies explicitly designed to compare immediate and delayed feedback in digital settings. Given the rapid advancements in educational technology and the increasing reliance on digital platforms, it is crucial to reassess whether earlier findings on

feedback timing still apply to contemporary computerized learning environments.

To address this gap, this thesis first presents a meta-analysis examining the effects of feedback timing in digital learning environments.

Recalling Previous Responses During Feedback Delivery: Another important question in feedback delivery is whether learners benefit from being reminded of their initial responses during feedback delivery. Several studies have suggested that the recall of an initial error was linked to improved error correction on subsequent tests (J. B. Knight et al., 2012; A. C. Butler, Fazio, et al., 2011; Iwaki et al., 2017; Yan et al., 2014). In these studies, participants first took an initial test on general knowledge questions, idioms, or word pairs and later received correct-answer feedback. On a delayed final test, they were asked to recall both their initial errors and the correct answers. Findings consistently showed that when individuals could recall their past errors, they were more likely to produce the correct response later. This suggests that memory for errors plays a role in strengthening associations with the correct answers, potentially facilitating learning.

Theoretical frameworks of memory provide possible explanations for this effect. One perspective is that recalling an error during feedback serves as a retrieval cue, linking the incorrect and correct responses in memory, which enhances retention (Barnes et al., 1959; Pyc et al., 2010). Another explanation is recursive reminding, a process in which recalling the discrepancy between an initial error and the correct answer strengthens both memories and reinforcing learning (Jacoby et al., 2015). However, not all studies support the notion that remembering past errors benefits learning. Some research has found no advantage or even a negative effect of recalling errors on later recall of correct responses (Loehr et al., 2020; Leggett et al., 2021; Metcalfe and Miele, 2014). Given these mixed findings, the role of recalling previous errors during feedback delivery remains a complex question, requiring further investigation to determine when it supports or hinders learning.

Building on the meta-analysis of feedback timing, this thesis includes a randomized controlled experiment conducted within the BNE medical training platform to examine the individual and interactive effects of feedback timing and response recall on learning outcomes in digital medical education.

4 Research Objectives and Thesis Structure

The primary objective of this dissertation is to optimize learning outcomes in medical education through learning analytics and randomized controlled experimentation in a digital learning platform.

At the start of this doctoral research, a comprehensive literature review was conducted to identify key challenges and research questions in optimizing digital learning, particularly in medical education. As summarized in previous sections, this review highlighted that while many findings from the learning sciences are well-established, their direct application to digital medical education requires further exploration. This need arises from the unique cognitive demands and constraints of medical learning, as well as the evolving landscape of learning environments, which continuously influence how learners process and retain information.

Building on this need, three fundamental points were identified as central to optimizing digital medical education in this doctoral research:

1. Measuring and tracking learning processes, which includes adapting a learning model to meet the specific needs of medical training data.
2. Optimizing the training difficulty levels in digital learning platforms.
3. Optimizing feedback delivery in digital learning platforms.

These areas form the foundation of this dissertation and guide the studies presented in the following chapters. Each following chapter explores a distinct aspect of these research areas, collectively contributing to the central question of this thesis: **How can training in digital learning systems be optimized to improve learning outcomes in medical education?**

The following section introduce the specific research questions addressed in each chapter.

Chapter 2 – Adapting a learning model to medical training data This chapter explores the application of a learning model—specifically, the Elo rating system— for estimating student ability level and question difficulty in a medical training platform, UNESS-BNE, introduced in the previous sections. The primary goal is to determine whether the model can effectively track and predict student knowledge in a complex, real-world educational setting, to document its limitations and applicability in this learning environment. To investigate this, learning analytics methods were applied, using historical learner records from BNE to estimate both student ability and question difficulty. These estimates were then used to evaluate the model's predictive accuracy on learners' final exam performance. The results show that the Elo rating system performs comparably to well-calibrated logistic models in predicting students' final exam outcomes, confirming its suitability for this dataset. This adapted model serves as the foundation for the next chapter, where it is used to further explore optimal training difficulty in medical education.

Chapter 3 – Investigating optimal training difficulty in medical learning This chapter examines the impact of training difficulty on learning outcomes within the UNESS-BNE platform. Building on prior research suggesting that an appropriate balance between challenge and achievability enhances learning, this study aims to determine whether an optimal level of training difficulty can be identified for medical students and whether it varies across medical specialties. To investigate this, a quasi-experimental design was used, analyzing BNE data where students were exposed to questions of random difficulty. Student ability and question difficulty were estimated using the adapted model from Chapter 2, which was then used to analyze the effect of mean relative training difficulty on final exam performance. This study contributes to the broader discussion on adaptive learning by providing empirical evidence on the role of difficulty calibration in a complex, knowledge-intensive domain.

Chapter 4 – A meta-analysis of feedback timing in digital learning environments While many studies have examined whether immediate or delayed feedback leads to better learning outcomes, the debate remains unresolved, as outlined in previous sections. Despite ongoing discussions on the optimal timing of feedback, a systematic examination specifically within digital learning environments is lacking. This chapter addresses this gap through a meta-analysis to determine the overall effect of feedback timing on learning outcomes in digital learning environments and identify factors that may explain inconsistencies across studies. This meta-analysis addresses three key questions: (i) What is the overall difference in learning outcomes between immediate and delayed feedback? (ii) How do different definitions of "immediate" and "delayed" feedback impact the reported effects on learning? (iii) What factors moderate the effects of feedback timing on learning outcomes? The results suggest that, overall, the timing of feedback does not have a significant effect on learning outcomes. However, moderator analyses highlight the influence of factors such as educational level, learning domain, post-test task type, and response time constraints, helping to explain some of the inconsistencies found in previous studies. The findings are then discussed in relation to their theoretical and practical implications for digital learning.

Chapter 5 – Investigating the effect of feedback timing and initial answer recall in medical learning This chapter examines the individual and interactive effects of feedback timing (immediate vs. delayed) and initial answer recall on learning outcomes in medical education. The central research question guiding this study is: How do feedback timing and initial answer recall interact to influence learning outcomes in higher-order learning tasks, such as those in medical education? To explore this, an experimental study was conducted within the BNE experimental module of UNESS. Participants completed multiple-choice questions under different feedback conditions, systematically varying both feedback timing and the inclusion of initial answer recall. Although the study was implemented at scale, the results did not reveal significant effects, likely due to limited participant engagement and reduced exposure to the experimental manipulation. However, data collection is ongoing, and future analyses are expected to provide more conclusive insights into the optimization of feedback strategies for complex

medical learning, where long-term retention and knowledge application are essential.

Chapter 6 – General Discussion This final chapter synthesizes the key findings of the dissertation, discusses its original contributions to the field, examines practical implications, and acknowledges its limitations.

Chapter 2

Adaptation of the Multi-Concept Multivariate Elo Rating System to Medical Students' Training Data

This section is based on the following published article:

Kandemir, E. N., Vie, J. J., Sanchez-Ayte, A., Palombi, O., & Ramus, F. (2024, March). *Adaptation of the Multi-Concept Multivariate Elo Rating System to Medical Students' Training Data.* In Proceedings of the 14th Learning Analytics and Knowledge Conference (pp. 123-133). https://doi.org/10.1145/3636555.3636858

Contents

ABSTRACT Accurate estimation of question difficulty and prediction of student performance play key roles in optimizing educational instruction and enhancing learning outcomes within digital learning platforms. The Elo rating system is widely recognized for its proficiency in predicting student performance by estimating both question difficulty and student ability while providing computational efficiency and real-time adaptivity. This paper presents an adaptation of a multi-concept variant of the Elo rating system to the data collected by a medical training platform—a platform characterized by a vast knowledge corpus, substantial inter-concept overlap, a huge question bank with significant sparsity in user-question interactions, and a highly diverse user population, presenting unique challenges. Our study is driven by two primary objectives: firstly, to comprehensively evaluate the Elo rating system's capabilities on this real-life data, and secondly, to tackle the issue of imprecise early-stage estimations when implementing the Elo rating system for online assessments. Our findings suggest that the Elo rating system exhibits comparable accuracy to the well-established logistic regression model in predicting final exam outcomes for users within our digital platform. Furthermore, results underscore that initializing Elo rating estimates with historical data remarkably reduces errors and enhances prediction accuracy, especially during the initial phases of student interactions.

1 Introduction

Over the last decade, as a result of the increasing use of educational technology involving significant amounts of data and learning analytics, personalized learning has gained increasing popularity. This has led a number of research groups to study the adaptation of personalized learning into educational technologies, leveraging insights from learning analytics (L.-K. Lee et al., 2020; Aljawarneh et al., 2021). Adaptive Learning Systems (ALSs) (J. Lee et al., 2008) achieve personalized learning experiences by using users' prior interactions and by adjusting the learning content to match individual preferences and requirements. Research consistently highlights the effectiveness of ALSs when compared to non-adaptive systems, resulting in improved learning outcomes (Lindsey et al., 2014; Tabibian et al., 2019; VanLehn, 2011; Ma et al., 2014) and a positive impact on student motivation (Abdi, Khosravi, Sadiq, and Gasevic, 2019), engagement (Papoušek and Pelánek, 2015), and comprehension (Alamri et al., 2020). Today, prominent ALS platforms like Duolingo, and ALEKS deliver high-quality learning materials and personalized instruction to millions of users worldwide.

With all the benefits listed above, this adaptive method in online education requires effectively monitoring users' learning paths, a procedure called knowledge tracing. This knowledge-tracing process involves creating learner models based on student performance and interactions with the system to represent their ability levels and the difficulty of educational materials.

1.1 Background

Various models have been designed to monitor and predict students' evolving knowledge levels over time (Brooks et al., 2017). These models can be broadly classified into four categories: Markov process models (Corbett et al., 1994; Gweon et al., 2015), logistic models (Choffin et al., 2019; Rasch, 1960; Linden et al., 2013; Lindsey et al., 2014), deep knowledge tracing models (Piech et al., 2015; Ghosh et al., 2020; Chan et al., 2021; Ruan et al., 2021; Shin et al., 2021), and rating systems (Abdi, Khosravi, Sadiq, and Gasevic, 2019). The first three classes of these models are well-established and already extensively documented in existing literature (Abdelrahman et al., 2023). Although they exhibit strong prediction capabilities when evaluating student performance, they are not without limitations, particularly when deploying them in online educational environments, where they often demand intricate parameter estimation and calibration procedures, typically relying on large datasets. This complexity can impede the development of adaptive systems, making them more challenging, time-consuming, and resource-intensive to create.

An intriguing alternative is the utilization of rating systems, offering a computationally more economical approach. Rating Systems, particularly the Elo Rating system (Elo et al., 1978), have been widely applied in educational technologies (Attali, 2014; Pelánek, 2016; Papoušek and Pelánek, 2017; Klinkenberg et al., 2011). Originally created for ranking chess players, the Elo rating system has been repurposed for educational settings, treating users and learning materials as opponents. In the educational context, it predicts ratings for users and questions, serving as an assessment of user ability and question difficulty.

In this framework, each user u is associated with a global ability parameter denoted as θ_u. Similarly, for each question i, there exists a question parameter θ_i reflecting the difficulty level of that question. The probability of a user u correctly attempting a multiple-choice question i, denoted as $\Pr(a_{ui} = 1|\theta_u, \theta_i)$, can be expressed as a logistic function of the difference between the user and question parameters:

$$\Pr(a_{ui} = 1|\theta_u, \theta_i) = \sigma(\theta_u - \theta_i) = \frac{1}{1 + e^{-(\theta_u - \theta_i)}}$$

After a user u attempts question i, both the user's ability and the question's difficulty undergo updates that are proportional to the difference between the estimated probability and the actual outcome. These updates are defined by the following formulas for question difficulty (θ_i) and for user ability (θ_u):

$$\begin{aligned} \theta_i &:= \theta_i + K(\Pr(a_{ui} = 1|\theta_u, \theta_i) - a_{ui}) \\ \theta_u &:= \theta_u + K(a_{ui} - \Pr(a_{ui} = 1|\theta_u, \theta_i)) \end{aligned} \tag{2.1}$$

Here, a_{ui} represents the actual outcome of the attempt of the user u on question i, and K is a constant value that determines the degree of update sensitivity based on the user's most recent attempt. The choice of the constant K in the update rule plays a pivotal role in shaping estimation dynamics. If K is set too low, the estimation process progresses too slowly, leading to prolonged uncertainty in skill assessment and potential failure to reach correct values. Conversely, if K is set too high, the system is unstable, heavily influenced by recent attempts, and thus provides erratic evaluations.

In light of this formulation, the classical Elo rating system in education reveals its iterative nature, refining user and question parameters after each interaction.

While the Elo rating system does not provide statistically guaranteed estimations, in contrast to well-calibrated logistic models, numerous studies have explored the accuracy of the Elo rating system and compared its performance to state-of-the-art models. For instance, Wauters, Desmet, and Van Den Noortgate (2012) found that the IRT-Rasch model version, proportion correct, and the Elo rating system, increasingly correlate with the true difficulty parameter as sample size increases. Another study (Papoušek, Pelánek, and Stanislav, 2014) compared Elo rating's question difficulty estimates with those obtained through the joint maximum likelihood method (JMLE) and observed nearly identical outcomes. Studies using simulated data (Antal, 2013; Pelánek, 2014; Pelánek, 2016) have also concluded that Elo rating systems perform similarly to the Rasch model, making them suitable for systems requiring real-time user knowledge adaptation without the need for extensive question pretesting on large sample sizes.

However, integrating the Elo rating system into a real-time educational context presents challenges, especially in the initial stages when student abilities and question difficulties are unknown and are set to zero. Given the iterative nature of the model, these initial estimates are assumed to gradually correct themselves with each attempt. While starting the model from scratch is standard practice, to produce reliable estimates the system requires a substantial number of responses, typically at least 100 attempts (Pelánek, 2016). Furthermore, uncertain

initial estimates can exert a lasting influence on subsequent updates, potentially resulting in persistent inaccuracies. This issue is especially pronounced for questions and users with a limited number of attempts within the educational platform, as they may have fewer opportunities for correction.

1.2 Goals of the present study

In this study, we have two primary objectives: firstly, to assess whether the Elo rating system meets the requirements of a complex real-life scenario with specific challenges, and secondly, to address the issue of imprecise early-stage estimations in the online application of the Elo rating system. To mitigate the uncertainty associated with initial Elo rating estimations, we used data collected in the previous year.

In brief, our learning platform is open to all French medical students throughout their studies to provide training for their medical studies (about 8600 students per year over 10 years, with one important national exam at the end of the 6^{th} year). The challenges raised by this particular learning context are multiple:

- The knowledge corpus is huge, as it encompasses *all medical knowledge* taught in French universities.
- This corpus is structured into knowledge components that are themselves very large: 362 distinct topics or subcategories (themes/areas) of medical knowledge, spread over 31 medical specialties.
- The corpus of questions is also huge (∼1,500,000 questions), such that a given student takes only a tiny fraction of available questions (usually drawn at random), almost never takes the same question twice, and such that a given question is only taken by a tiny fraction of students (outside exams). Thus, the matrix is extremely large and sparse.
- Use of the training platform is optional, with some students using it intensively on a daily basis, and others doing most of their training outside the platform.
- Students are based in 42 universities which cover the same curriculum from 1^{st} to 6^{th} year, but with different material and in a different order.

Despite these difficulties, the fact that all medical students from all universities can train and take exams on the same platform should make it possible and desirable to model their progress and use this modeling to provide them with an adaptive training program. Yet the first step is to show that a sufficiently reliable modeling of their progress is possible under the present conditions.

Thus, the main research question guiding this study is to examine the boundaries of the Elo rating system within a particularly challenging real-life scenario that encompasses various complexities. This evaluation involves a comparison of its accuracy against the widely accepted

logistic regression model. Additionally, we seek to explore the potential advantages of initializing the Elo rating system with results obtained from logistic regression applied to data from preceding years.

2 Methods

This study employed an observational research design, leveraging ecological data from the existing BNE (*Banque nationale d'entraînement*) digital learning platform.

2.1 BNE Platform

The BNE digital learning system serves as an online platform extensively utilized by over 8,800 medical students across all French universities annually. This platform is used by all medical faculties to administer exams. Exam questions are then added to the question bank, together with additional questions designed by professors for training purposes. This platform therefore holds a very large set of multiple-choice questions covering 31 medical specialties that are made available to students for training. For medical students, the platform is a valuable resource to prepare for the ECN (*Épreuves Classantes Nationales*) final exam. This final exam usually takes place in June of the sixth academic year and significantly influences the choice of students' medical specialization.

To enhance the pedagogical engagement of medical students on the platform, a notable feature allows learners to tailor their training experience. They can choose from various question types and medical specialties, enabling them to simulate and practice for a wide range of medical exams according to their preferences and needs.

2.2 BNE data set Overview

Within this section, we describe the BNE data set for the educational year 2020-2021, representing the most recent accessible data sourced from the BNE platform during our analysis. Additionally, we describe the usage patterns observed on the platform during this year, providing valuable insights into its structure and functionality.

From the BNE platform, direct access to the official ECN exam is unavailable. However, we do have access to a mock exam, typically conducted in mid-March (specifically on March 15^{th}, 16^{th}, and 17^{th}, 2021, for the 2020-2021 academic year). This mock exam closely mimics the format of the actual ECN final exam. For the purpose of testing student prediction models on our complex data set, we focused our analysis on the educational year of 2020-2021, specifically targeting 6^{th}-year users who participated in the mock final ECN exam. This selection aligns with the core objective of our study, which is to assess the models' performance through external validation using the mock final exam.

In the following sections, we will describe the data related to the six-month training period spanning from September 16, 2020 to March 14, 2021, leading up to the mock final exam, distinct from the data associated with the mock exam itself, on March 15^{th}, 16^{th}, and 17^{th}, 2021.

Table 2.1: **BNE Data Set Summary**

Data	Period	Users	Questions	Specialties	Specialties per Question	Attempts	Sparsity (User, Question)	Attempts per User
Training Period Data Set	**16.09.2020–14.03.2021**	8,616	357,317	31	1.58	26,772,424	0.99	1.05
Mock Final Exam Data Set	**15–17.03.2021**	8,616	372	28	1.71	3,172,546	0.01	1

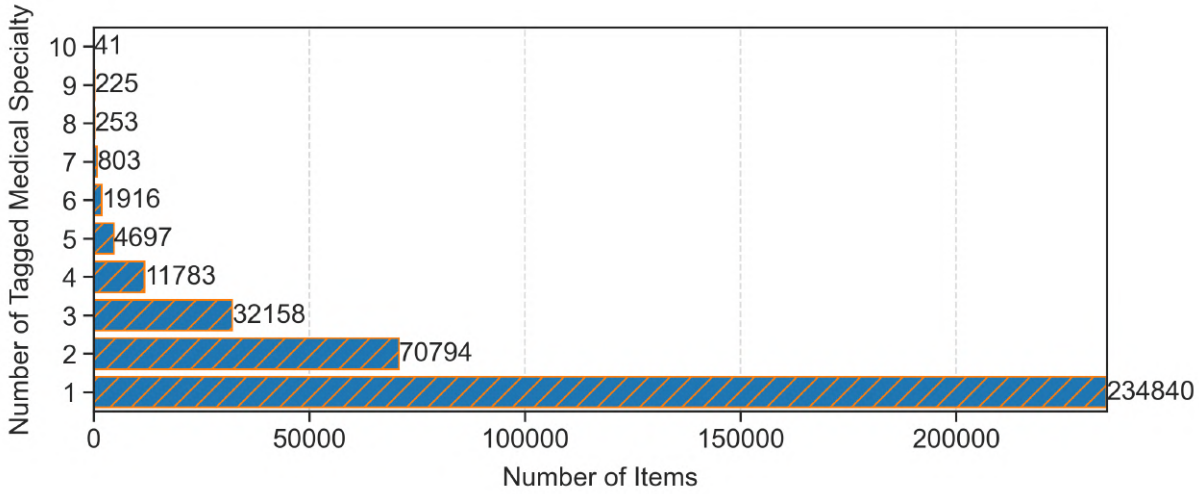


Figure 2.1: Number of questions that require knowledge on any given number of medical specialties.
The questions exhibit a spectrum of dependence on medical specialty knowledge for their solution. While a considerable portion of questions rely on ability in a single medical specialty, many questions require knowledge spanning multiple specialties.

2.2.1 Training Period data set

Table 2.1 offers a comprehensive overview of our training period data set's key characteristics, including the total number of *users* (8,616), *questions* (357,317), *medical specialties* (31), and *attempts* (26,772,424).

Additionally, the *specialty per question* variable indicates the average number of specialty tags associated with each question. Here, each of the 31 medical specialties serves as a distinct knowledge component (KC). These knowledge components are much larger than is usually defined in the literature. In this context, when a question i is tagged with a specific specialty s, the likelihood of correctly answering question i hinges upon the user's specialty-specific ability s. Conversely, we assess a user's ability in specialty s by evaluating their ability to successfully tackle questions tagged with the same specialty s. The table reveals that the average number of specialties associated with each question (1.58) is greater than 1, underscoring that certain questions require knowledge spanning multiple medical specialties for accurate responses. In

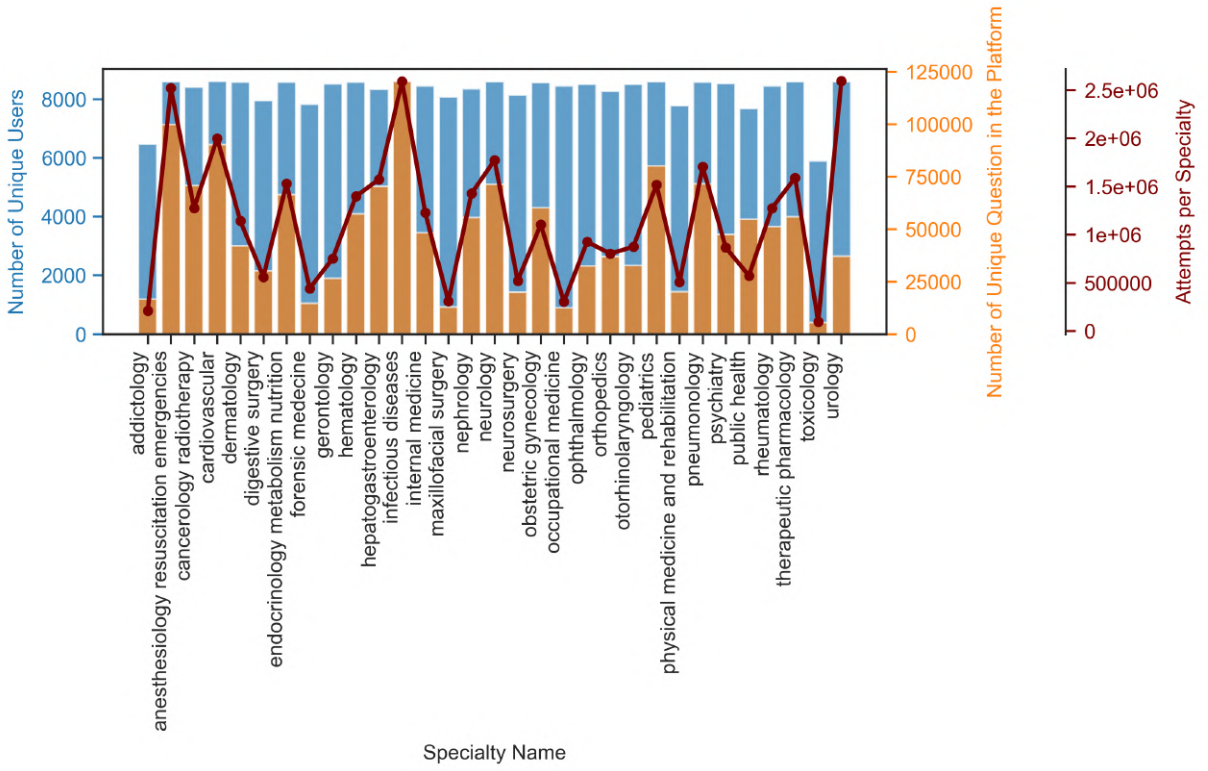


Figure 2.2: Overview of the use of the BNE Platform during the 2020-2021 educational year. The blue bars represent the count of unique users per medical specialty in the data set. The orange bars represent the count of unique questions available in the platform for each specialty. In addition, the overlaid line plot illustrates 'Attempts per Specialty,' the total number of user attempts on questions within each specialty during the educational year 2020-2021.

this context, our data set can be characterized as a multi-knowledge component data set (or multi-specialty data set in the specific context of BNE). For a more in-depth exploration of this distribution, Figure 2.1 provides a detailed breakdown of the count of questions requiring varying numbers of specialties.

The user-question sparsity in Table 2.1 indicates the proportion of missing values in the user-question interaction matrix. The table shows that our data set is substantial and exhibits significant sparsity (*Sparsity (user, question)* = 0.99). There are vastly more available questions than any user can take, and different users will take different questions (usually by random draw).

Lastly, the *Attempts per User* variable unveils the average number of attempts made by the same user on individual questions. This indicates whether students frequently revisit questions they have previously encountered. With a value of 1.05, it is evident that during the training period, students almost never re-attempt questions they have already attempted.

Figure 2.2 provides a detailed description of the platform's usage patterns across 31 available medical specialties during the 2020-2021 educational year. As mentioned earlier, the platform provides users with the option to create their own training sessions by selecting specific medical specialties and question types they want to study. This results in varying levels of popularity

among the specialties. The figure reveals that while most students engage with questions from all specialties, there is considerable variability in both the quantity of available questions and the number of questions taken within each specialty.

2.2.2 Mock Final Exam data set

Table 2.1 also provides a comprehensive overview of the nature of the mock ECN final exam for the educational year 2020-2021, held on the 15$^{\text{th}}$, 16$^{\text{th}}$, and 17$^{\text{th}}$ of March 2021. The Mock final exam data set encompasses data from 8,616 users who took 372 questions spanning 28 distinct medical specialties (addictology, orthopedics, and toxicology were not included in the mock exam). Questions in this data set were associated with a mean number of 1.71 specialties, reflecting greater multidisciplinarity of questions than in the training period data set. The mock final exam data set recorded a total of 3,172,546 user interactions, highlighting a high level of user engagement, with a minimal sparsity value of 0.01, implying that almost all users attempted every question during the final exam. Additionally, the *Attempts per User* value of 1 indicates that users made only a single attempt at each question.

Here, it's worth noting that all the mock exam questions were entirely new, distinct from those encountered during the training period. Therefore, the data from the mock final exam does not provide a direct test of the knowledge of specific questions taken in the training period, but rather a test of students' ability to generalize what they have learned during courses and training to new questions in the same medical specialties, mirroring the format of the official ECN exam, which emphasizes generalization rather than memorization of specific knowledge.

2.3 Elo Rating System: Model Extensions for Adaptation to the BNE data set

The standard iterative formulation of the Elo rating system, which computes user and question-related factors, has been previously described in the related literature (Attali, 2014; Pelánek, 2016; Papoušek and Pelánek, 2017; Klinkenberg et al., 2011; Antal, 2013; Pelánek, 2014; Papoušek, Pelánek, and Stanislav, 2014). To optimize its adaptability for educational contexts, the Elo rating system has undergone numerous extensions. In this section, we indicate the specific modifications we have applied to tailor the model to our BNE data set.

2.3.1 Incorporating the guessing behavior into the Elo rating system

In numerous studies employing the Elo rating system as a prediction model for multiple-choice questions, researchers take into account the guessing rate when calculating the probability of correctness (Papoušek, Pelánek, and Stanislav, 2014; Pelánek, 2016).

In such instances, the probability of a user u attempting a multiple-choice question i with n_{opt} choices correctly, denoted as $\Pr(\text{correct} \,|\, \theta_u, \theta_i)$, can be expressed as follows:

$$\Pr(a_{ui} = 1 \,|\, \theta_u, \theta_i) = \Pr(\text{guessing} \,|\, n_{\text{opt}}) + \frac{1 - \Pr(\text{guessing} \,|\, n_{\text{opt}})}{1 + e^{-(\theta_u - \theta_i)}}$$

Within the BNE question pool, questions are divided into two main types: single- and multiple-answer questions. For single-selection questions (unique answer questions), $P(\text{guessing}|n_{\text{opt}})$ is straightforward, equating to $1/n_{\text{opt}}$. For multiple choice questions, $P(\text{guessing}|n_{\text{opt}})$ is the inverse of the sum of the possible ways to select any number of answers k from the available options:

$$\Pr(\text{guessing} \mid n_{\text{opt}}) = \frac{1}{\sum_{k=1}^{n_{\text{opt}}} \binom{n_{\text{opt}}}{k}}$$

2.3.2 Decreasing Uncertainty

The dynamics of the Elo rating system in educational settings are crucial for accurately assessing the skills and abilities of students and questions. The challenge lies in managing evolving uncertainties, which are inherently dynamic. When new students or questions are introduced to the platform, our information on their true abilities or difficulties is limited, meaning high uncertainty. Consequently, during this initial phase, it is essential for the model to make significant updates to its estimates. As more data accumulates, students engage in multiple attempts, and questions are extensively attempted by a set of students, and the model should naturally become more certain about its estimation of ability levels or difficulty levels. In such cases, the model should reduce the update parameter as confidence in the estimates grows.

In order to meet this challenge, recent applications of the Elo rating system in educational contexts (Abdi, Khosravi, Sadiq, and Gasevic, 2019) have replaced the fixed constant K in Equation 3.2 with a dynamic uncertainty function. This function, denoted as $U(n)$, is defined as:

$$U(n) = \frac{a}{1 + bn}$$

where a is the constant hyper-parameter determining the starting value; b is the constant hyper-parameter determining the slope of changes; n is the number of prior attempts of the user or question parameter.

The exact parameter values, as highlighted by Pelánek (2016), carry relatively less weight, as different choices for a and b tend to yield remarkably similar outcomes. In our case, we set $a = 1$ and $b = 0.5$ for both question and user attempts. These values were determined through an optimization process using grid search. However, it is important to mention that our model consistently delivered stable performance, and the precise choice of parameter values had only a negligible effect on the results.

Moreover, in keeping with the central aim of learner models, which is to effectively track shifts in user abilities, we have introduced a lower bound for the uncertainty function applied to user ability. By incorporating this lower bound of 0.03 into the user uncertainty function, we ensured that user ability updates persist even after a considerable number of attempts. With our current values of a and b, this lower bound applies after 65 attempts.

2.3.3 Multi-tag Knowledge Component Extension

As previously described, in our BNE data one question may be tagged by multiple specialties. To account for the ability of users in each of the tagged medical specialties separately we used

the multi-concept extended version of the Elo rating system introduced by Abdi, Khosravi, Sadiq, and Gasevic (2019). The difference is that, instead of having only one global user ability parameter θ_u, we estimated user ability θ_{us} for each specialty s.It is important to note that, given the absence of information regarding the relative importance of tagged specialties for each question in the data, we adopted a straightforward approach as outlined in Abdi, Khosravi, Sadiq, and Gasevic (2019). Specifically, we computed the mean ability λ_{ui} of student u on question i by averaging user u's abilities across all medical specialties $s_1, \ldots, s_\delta$ tagging question i, assigning equal weight to each specialty in this calculation.

$$\lambda_{ui} = \frac{1}{\delta} \sum_{k=1}^{\delta} \theta_{us_k}$$

Furthermore, not all specialties may have the same average difficulty level. In order to alleviate this, we define and estimate distinct parameters for question difficulty (d_i) and specialty difficulty (θ_s). We denote by μ_i the sum of question difficulty d_i and the average of difficulties of all skills $s_1, \ldots, s_\delta$ involved in question i:

$$\mu_i = d_i + \frac{1}{\delta} \sum_{k=1}^{\delta} \theta_{s_k}$$

Thus the probability of answering correctly becomes:

$$\Pr(a_{ui} = 1 \lambda_{ui}, \mu_i) = p(\lambda_{ui}, \mu_i) \triangleq \sigma(\lambda_{ui} - \mu_i)$$

The update of the question difficulty d_i remains the same. However, now the update on the user skill parameters θ_{us} occurs on each tagged specialty separately, and the update for specialty difficulty θ_s follows a similar pattern as the updates for item difficulty:

$$\begin{aligned} d_i &:= d_i + U(n) \left(p(\lambda_{ui}, \mu_i) - a_{ui}\right) \\ \theta_{us} &:= \theta_{us} + U(n) \left(a_{ui} - p(\theta_{us}, d_i + \theta_s)\right) \\ \theta_s &:= \theta_s + U(n) \left(p(\theta_{us}, d_i + \theta_s) - a_{ui}\right). \end{aligned}$$

It's important to note that while updating θ_{us} and θ_s, the prediction formula operates at the specialty level for each tagged specialty, just like in Abdi, Khosravi, Sadiq, and Gasevic (2019), although the d_i update is based on question-level prediction.

By utilizing the Elo rating system along with the aforementioned extensions, it becomes possible to estimate three critical aspects: user ability in each specialty, questions' individual difficulty, and specialties' global difficulty.

2.4 Data Preparation Process

Before starting to train the Elo rating system and Logistic regression over the 2020-2021 data set, we performed a series of pre-processing steps on the combined data from the training period data set and the mock ECN final exam data set. These steps were carried out in the following order:

- Removal of duplicated rows: 267 rows out of 29,900,533 were removed.
- Exclusion of questions without any tagged medical specialty: None removed (the data extraction process was already limited to questions with tagged specialties), but 30% lacked specialty tags initially.
- Exclusion of questions that are neither unique nor multiple-choice questions (open-answer questions): None removed.
- Binarization of question ratings (BNE has a more sophisticated rating scheme depending on the number of correct and incorrect answers ticked).
- Removal of users with fewer than 100 interactions: No questions or students were removed during this step. Since all students in the dataset had taken the ECN mock exam, they all had at least 100 attempts.
- Removal of questions with fewer than 100 interactions: 79.11% of the unique questions were removed.

As a result, our training period data set now consists of 22,294,780 attempts, made by 8,616 distinct users to 74,704 unique questions across 31 medical specialties. The mock ECN final exam data set includes 3,172,546 attempts. Within that data set, there are 372 unique questions taken by 8,616 users across 28 distinct medical specialties.

Figure 2.3 offers a visual depiction of the distribution of answers across students, questions, and specialties in both the training period and the mock final exam data sets after the pre-processing.

2.5 Information Encoding & Initialization of Elo Ratings System via Logistic Regression Outputs

As previously mentioned, in the Elo rating system, initial estimates for both questions and users are typically set to 0, which can lead to high uncertainty. To address this, an alternative approach is to use the logistic regression outcomes of the previous year's data as informed initial values for initializing Elo ratings, rather than starting from scratch. With this approach, the Elo rating system is anticipated to converge faster and more accurately, providing a "head start" that conserves computational resources and leads to more precise estimates.

To prepare the extensive BNE dataset for logistic regression modeling, we employed a one-hot vector encoding method inspired by Vie et al. (2019). This technique transformed each attempt in the original data set into a sparse vector containing all relevant information. In our adaptation of this approach, we aimed to closely align our logistic regression model with the principles of Elo rating estimations, while also incorporating all the aforementioned extensions we applied for the Elo rating system. To achieve this, we included the one-hot encoding of user-specialties interaction, question, and specialty for each attempt. With this approach, the logistic regression was able to estimate users' ability in each specialty, the difficulty of individual questions, and the overall difficulty of each specialty.

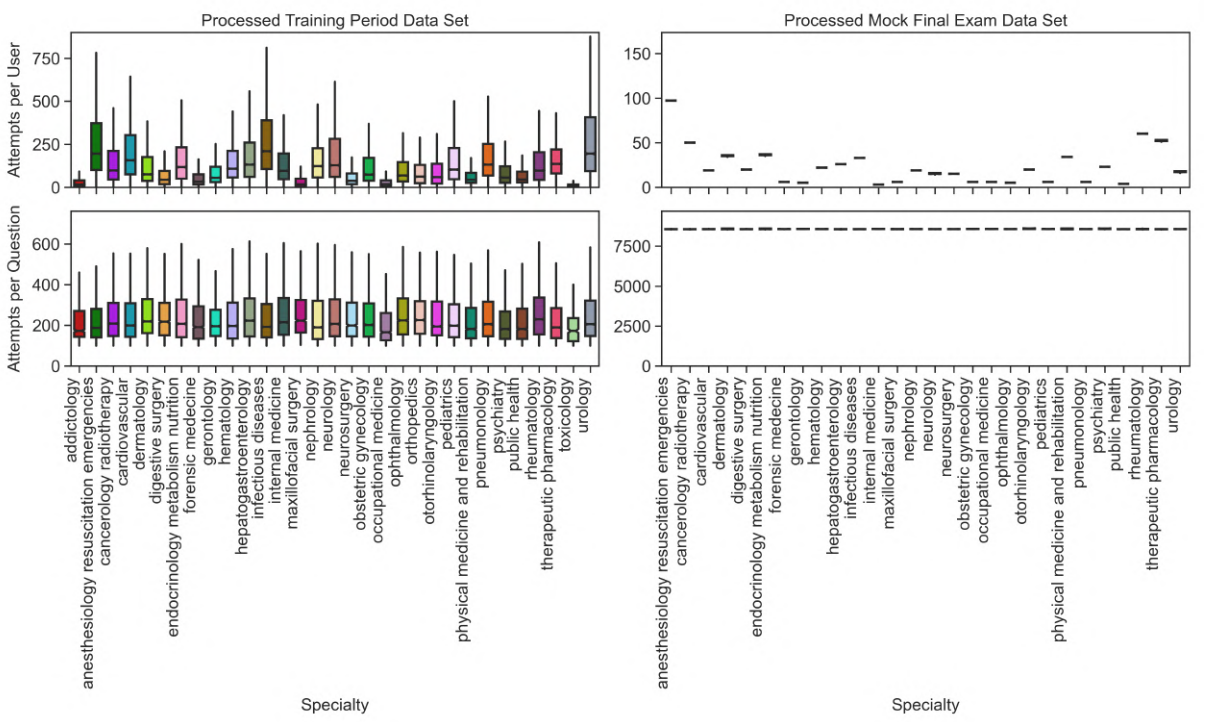


Figure 2.3: Number of Attempts by Each User and to Each Question across the 31 Medical Specialties.
The top left box plot shows the distributions of the number of attempts by each user across the 31 specialties. The bottom left box plot depicts the number of attempts to questions in each specialty. During the mock exam, all students took identical questions, resulting in quasi-uniform numbers of attempts given by students and received by questions (top and bottom right plots).

To implement the initialization approach, an essential step involves comparing the logistic regression model against the Elo rating system, utilizing data from the same year. This step aimed to ensure that, before employing the logistic regression model on the previous year's data and utilizing its outcomes for initialization purposes, the model aligned with the Elo framework, generating compatible and consistent results.

Subsequently, we applied the logistic regression model to the data from the 2019-2020 educational year, utilizing the outcome estimates as initial values for the Elo rating process applied to the 2020-2021 data. Our data set for the academic year 2019-2020 (spanning from September 15, 2019, to March 1, 2020) includes 400,774 distinct questions and 25,978 unique users. However, after filtering data to retain questions and users with enough attempts (cf. above), only 47,579 questions and 8,239 users were shared between 2019-2020 and 2020-2021.

As a result, we initialized the question difficulty and student ability values for the 2020-2021 academic year using the estimates obtained from the logistic regression model applied to the 2019-2020 data, whenever these values were available. In cases where values were not present in the 2019-2020 data set for a particular question or student, we initialized their 2020-2021 values to zero.

In order to allow the uncertainty function to be able to further update those initialized values, without entirely destabilizing the estimations, we set the initial number of previous attempts to 50, which seemed a reasonable compromise between the actual number of attempts (>100 which would make any update negligible) and 0 (which would underweight the previous history).

2.6 Performance Evaluation Metrics

In our performance evaluation, we examined the effectiveness of two variants of the Elo rating system on the training data set. One variant initialized all ability and difficulty values to 0, while the other initialized values based on logistic regression from the previous year. We also compared these Elo variants with logistic regression on the entire training data set. We used several key statistics, including Area Under the Curve (AUC), Root Mean Squared Error (RMSE), and Accuracy (ACC), to assess the prediction performance of these models first on the training period and second on the mock exam.

First, to comprehensively evaluate the prediction capabilities of the Elo rating system throughout its iterations, mirroring its real-world use within the platform, and understand the impact of initializing estimates via logistic regression, we compared these three models during the training period. We calculated the AUC, ACC, and RMSE scores for each training period day from September 16, 2020, to March 14, 2021, providing insights into how these models adapted and remained robust over time.

Subsequently, we turned our attention to evaluating these models' ability to predict performance on the mock exam using the same metrics: AUC, RMSE, and ACC. However, the mock exam posed a unique challenge as it featured entirely new questions that were not part of the training period. To address this challenge, we needed to estimate the difficulty of these mock exam questions. Our approach involved combining the entire training data set with a random selection of 60% of user data from the mock exam data set as the train set while designating

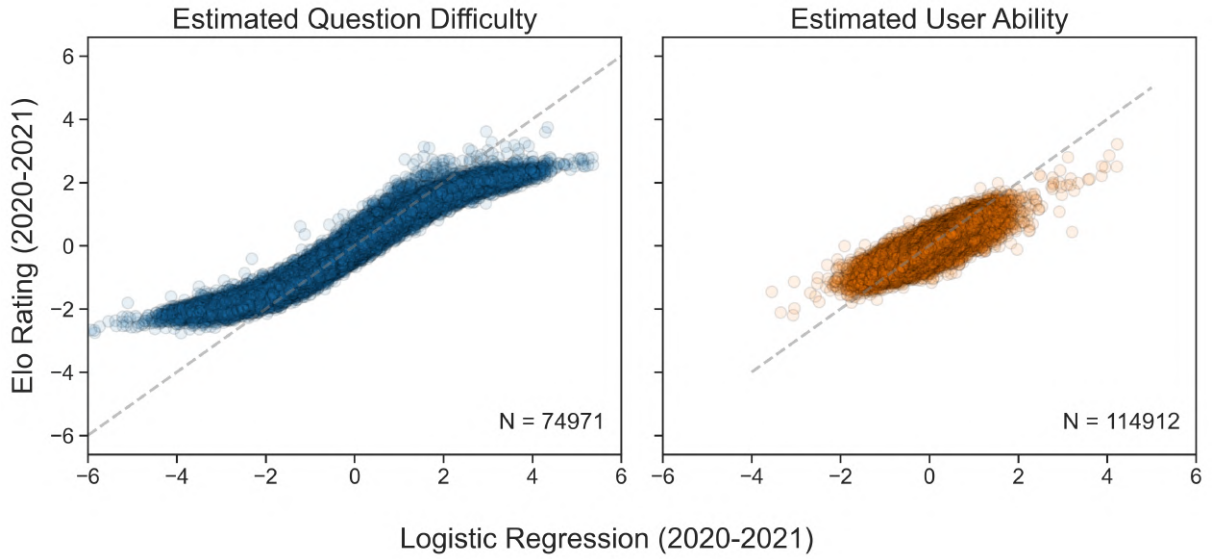


Figure 2.4: Comparing Logistic Regression and Elo Rating System outcomes for question difficulty (left) and user ability across 31 specialties (right) in the same 2020-2021 dataset. Scatter plots illustrate the alignment, with $y = x$ lines for reference. The left plot displays Logistic Regression difficulty estimates on the x-axis and Elo Rating System estimates on the y-axis. On the right, the plot contrasts user ability estimates, with Logistic Regression on the x-axis and Elo Rating System on the y-axis. Sample sizes (N) are included in each plot.

the remaining 40% of user data from the mock exam data set as the test set. This allowed us to create a training set that encompassed all attempts, including those from the mock exam, for 60% of users. For the remaining 40% of users, we included only their attempts from the training period in the train set. By doing so, when we ran the learner models on the training set, we obtained estimates of ability for all students and difficulty for all questions, which in turn enabled us to measure the models' prediction ability on the mock exam.

Following this data division process, the training subset comprised a substantial 24,152,933 entries, involving 8,616 unique users and 74,971 unique questions. In parallel, the test subset consisted of 1,268,752 entries, encompassing 3,447 unique users and 372 unique questions.

To assess the prediction ability of the models on the mock exam, after executing the models on the specified training set and obtaining difficulty estimates for all questions and ability estimates for all students, we evaluated its prediction performance on the test set. This evaluation capitalized on the stabilized estimates derived from the comprehensive training data set.

3 Results

3.1 Correlation between the Estimates

Figure 2.4 illustrates a notable positive correlation between the estimates of question difficulty ($r = 0.97$) and user ability ($r = 0.86$) derived from the Elo rating system and logistic regression models for the same-year data. This strong positive correlation clearly indicates that both models converged toward similar final estimations regarding question difficulty and user ability.

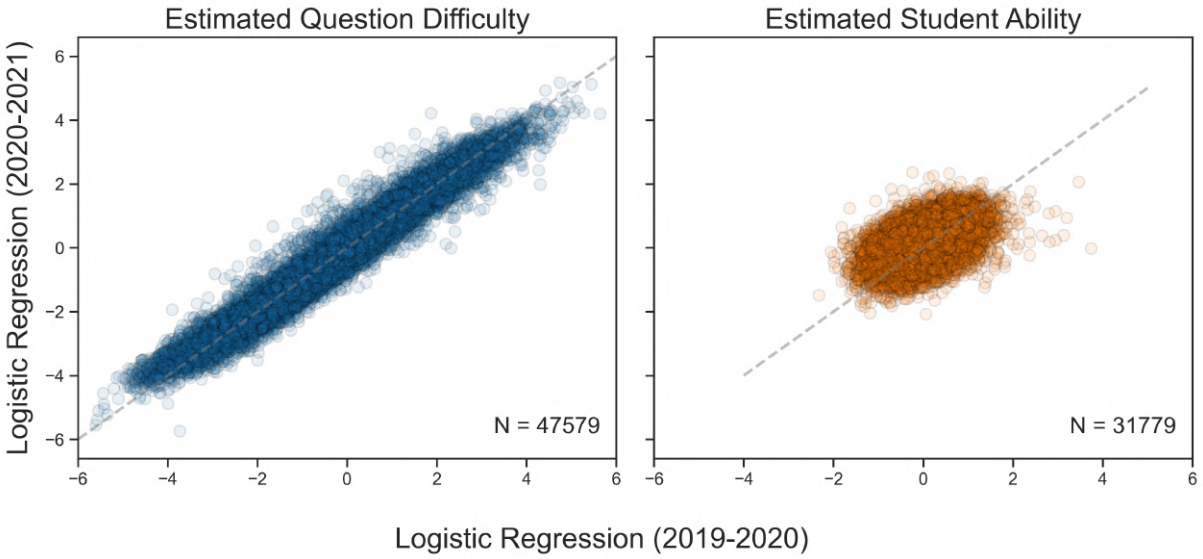


Figure 2.5: Comparing Logistic Regression Outcomes.
Estimated question difficulty (left) and user ability on each of 31 specialties (right) in the two successive education years (2019-2020 and 2020-2021) using the Logistic Regression model. $y = x$ lines are given for reference.

Figure 2.5 shows the question difficulty and student ability estimations using the logistic regression in the two successive years. Specifically, for shared questions, we observe a robust positive correlation of 0.98, indicating that the difficulty levels of these questions remain relatively consistent over time. However, when it comes to students' abilities, the correlation, albeit positive at 0.54, is not as strong. This suggests that students' abilities in various specialties have undergone some changes over the years, as we anticipated.

3.2 Prediction Accuracy

Figure 2.6 presents a visual representation of the RMSE and AUC scores over 180 consecutive training days for each model. Findings indicate a substantial prediction accuracy advantage at the beginning of the training year when initializing the ability and difficulty values based on the data from the previous year. This advantage is reflected in an initial boost in average accuracy (+0.016 AUC) and a reduction in the average error (-0.008 RMSE) during the initial 30 days of training. However, it's worth noting that the initial disparity between the two model versions diminishes rapidly and becomes less than 1 point within a few days. By the end of the training period, the advantage of initializing with historical data becomes nearly negligible, with only a marginal improvement in average accuracy (+0.002 AUC) and a minimal reduction in average error (-0.002 RMSE) observed during the last 30 days of training.

Table 2.2 shows the prediction performance on the mock exam for the three models: Elo rating system initialized at 0, Elo rating system initialized with historical data, and logistic regression. These results reveal that the three models show highly similar prediction accuracy, with a slightly better performance on the Elo rating system initiated with historical data over other models.

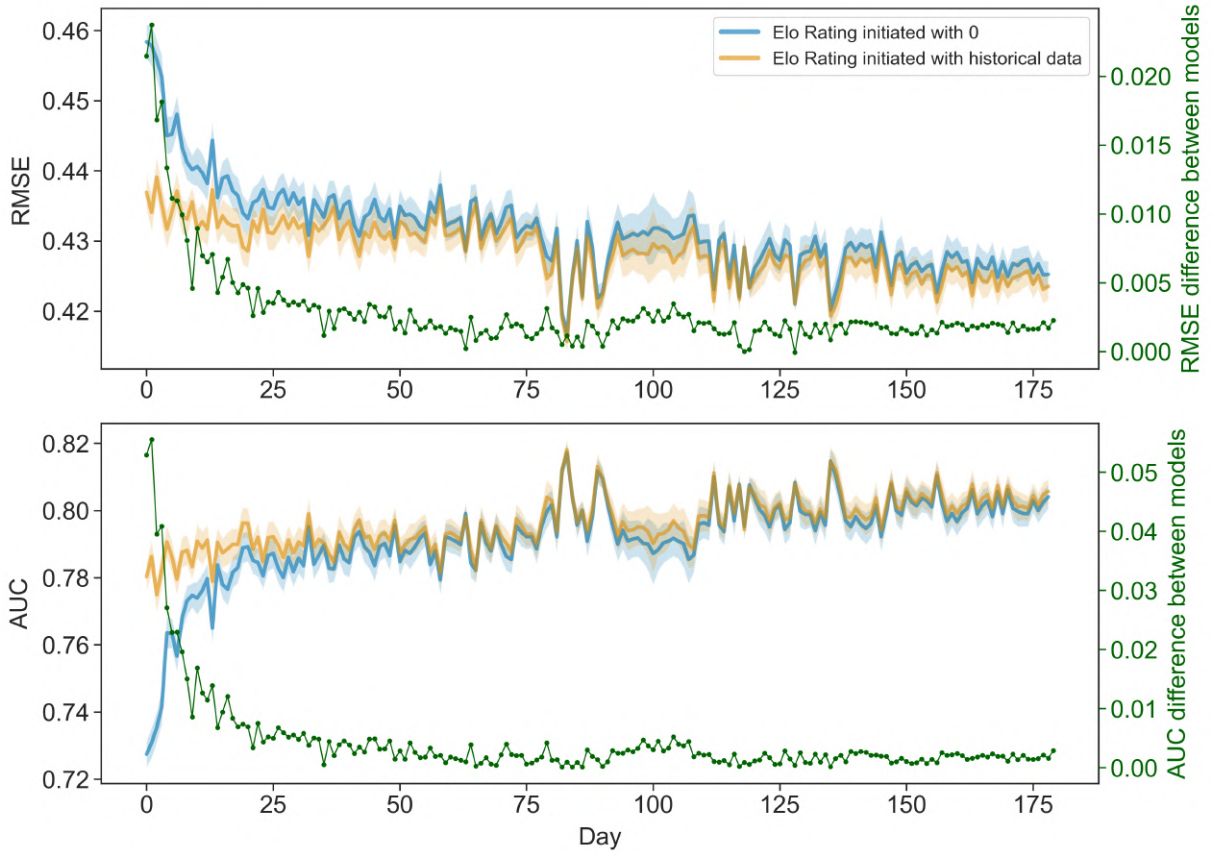


Figure 2.6: Comparing Models' Prediction Performance During Training.
RMSE (top) and AUC (bottom) values as a function of training days, for two versions of the Elo rating system. Shaded regions around the mean lines represent the 95% confidence intervals calculated using the standard error. A secondary y-axis on the right side illustrates the absolute difference between the two models with the green line plot.

Table 2.2: Comparing Models' Prediction Performance on Mock Exam

Model	RMSE (↓)	AUC (↑)	ACC (↑)	Precision (↑)	Recall (↑)	F1 (↑)
Elo rating system initialized at 0	0.419	0.812	0.737	0.717	0.685	0.701
Elo rating system initialized with historical data	0.418	0.813	0.738	0.720	0.684	0.702
Logistic regression	0.419	0.811	0.736	0.714	0.689	0.701

4 Discussion

This research is an initial step in integrating the multi-concept Elo rating system into our medical training platform in order to achieve real-time estimates of user performance. The results demonstrated the Elo rating system's comparable prediction power to the logistic regression models, confirming its suitability for this specific data set.

The Elo rating system offers several significant advantages in the context of our medical training platform. Firstly, the multi-concept Elo rating system excels in estimating concept-level competencies, which is crucial for tailoring adaptive learning experiences in our data in which questions mostly require knowledge from multiple medical specialties. Secondly, it stands out as a computationally efficient and cost-effective option especially in real-time estimations, compared to logistic regression models which require processing a vast amount of previously collected data.

Given our primary objective of identifying the most effective prediction model for online applications in our platform, and considering the challenges associated with logistic regression in terms of online adaptability, our focus naturally shifted towards a more detailed comparison of the two versions of the Elo rating system: one starting from scratch at the beginning of the year, and one with difficulty and ability values initialized based on the previous year's data.

Our findings underscore that while the overall performance in predicting mock exam results remains very similar regardless of the initialization approach, a distinct advantage emerges for initialization based on historical data, particularly during the initial phases of iteration. This may be important in scenarios that demand accurate estimations from the outset, such as real-time or online applications. This approach of initializing the model with historical data enhances the model's ability to produce quicker and more precise estimates, thereby enhancing the reliability of personalized learning environments utilizing Elo rating systems.

4.1 Unique Characteristics of the Data set

Our data set has very broad knowledge components, made of 31 distinct medical specialties, each of which is a huge corpus of information. Moreover, this platform prioritizes comprehension over memorization: questions are hardly ever repeated twice. Students have to generalize their knowledge while attempting the questions. This departure from purely memory-based learning provides an excellent chance to assess the model's performance in dealing with non-repeated inputs. While this limited exposure to questions challenges standard prediction models used for repeated question iterations, it also allows us to evaluate the model's flexibility in settings where students rely on wider conceptual knowledge rather than memorized responses. Additionally, our platform allows students to personalize their own training experiences. They have the option of selecting the medical specialty in which they want to train and the type of questions, resulting in unique interaction patterns for each student. This adds another layer of complexity to our data set. In addition to these complexities, we also lacked control over learning occurring outside the platform.

Despite these challenges, it is remarkable that the Elo rating system has achieved significant

prediction power for assessing the accuracy of students' future responses, with about 73.7% accuracy and 0.81 AUC. This demonstrates the model's adaptability and endurance in circumstances that deviate from standard educational data. In addition, the Elo rating system has shown good online prediction accuracy during training, right from the start when initializing with historical data, and after about 15-20 days when starting from scratch.

With the overarching goal of transforming our medical training platform into a personalized and adaptive format through knowledge tracing methods, we carefully considered the data set's characteristics. The Elo rating system stood out as a prime choice due to its simplicity, rapid parameter estimation, and real-time knowledge assessment capabilities. Its straightforwardness, coupled with widespread use in applications like online games and chess, makes it easily explainable compared to more complex models. An exemplar of transparency in a multivariate Elo version is evident in the study by Abdi, Khosravi, Sadiq, and Gasevic (2019). The study demonstrates the feasibility of making the algorithm transparent to students, a practice that not only heightened their motivation to engage with the platform but also enhanced their trust in the recommendations provided.

The implemented extensions further enhance adaptability to our data set's unique characteristics, offering optimization along with advantages in suitability and transparency. Notably, the multi-concept Elo rating system, in contrast to its single-concept version, acknowledging the non-transitive nature of skills, provides a realistic representation of learners' capabilities, crucial for accommodating interdependencies within medical specialties, potentially involving prerequisites. Thus, the multi-concept Elo rating system emerges as a fitting and transparent knowledge-tracing method for our complex data set.

4.2 Limitations and Further Work

One major limitation is that our tested models (logistic regression and Elo rating system) do not consider the natural forgetting of knowledge over time, which is well-documented in human memory research dating back to Ebbinghaus in 1885 (Ebbinghaus, 2013). Incorporating learning and forgetting curves into prediction models, as shown in research like DAS3H (Choffin et al., 2019) and MV-Glicko (Abdi, Khosravi, and Sadiq, 2021), which builds upon the multivariate-Elo rating system (Abdi, Khosravi, Sadiq, and Gasevic, 2019), can improve the models' prediction accuracy.

Our study also suggests several promising future research directions. One area of focus is improving learning models to better assess question difficulty and student ability, especially for topics requiring knowledge of multiple concepts. This involves determining the importance of each concept in problem-solving and investigating how these concepts interact during the learning process. Additionally, comparing the Elo rating system with the Glicko rating model within our data set could provide insights into the role of learning and forgetting curves in student performance estimation.

Beyond the aforementioned research areas, a critical future direction involves implementing the Elo rating system for online recommendations regarding specialties and question difficulty. However, this endeavor presents challenges in platforms where questions often have multiple

specialty tags. As underscored by the research findings of Nicholas J. Cepeda et al. (2008), the premature revisiting of knowledge can exert detrimental effects on long-term memory. Consequently, during the recommendation phase, it becomes imperative to not only select questions aligned with the student's current needs but also to ensure that these questions do not encompass specialties that may not be relevant to the student's current stage of learning. To address this, we can consider a new approach that calculates students' abilities based on combinations of specialties, rather than individual ones. This new strategy could substantially enhance the model's efficacy when suggesting questions that align with students' learning requirements and are pertinent to their current learning stage. Such an approach would ensure that students are presented with a tailored set of questions that optimally support their progress while avoiding the unnecessary revisiting of topics that might hinder long-term retention—a crucial consideration for the success of an adaptive learning platform.

5 Conclusion

In conclusion, our study underscores the remarkable adaptability of the Elo rating system to the intricate challenges posed by a large, sparse, and multifaceted data set, where questions are tagged with multiple knowledge components. The Elo rating system, along with its enhanced version that leverages historical data for initial estimations, has exhibited a commendable level of prediction accuracy.

These results offer reassurance regarding the Elo rating system's robustness and versatility, emphasizing its capability to provide reasonable predictive value even in complex situations. This insight is crucial for the broader learning analytics community, providing confidence in the effectiveness of the Elo rating system as a predictive model in educational settings. Overall, the study contributes to the ongoing discourse on learning analytics methodologies, offering practical insights and encouraging further exploration of the Elo rating system's applicability in diverse learning scenarios.

Chapter 3

Investigating the Influence of Training Difficulty on the Learning Outcomes of Medical Students

This section constitutes the following manuscript:

Kandemir, E. N., Vie, J. J., Sanchez-Ayte, A., Palombi, O., & Ramus, F. (2025). *Investigating the Influence of Training Difficulty on the Learning Outcomes of Medical Students. (under review)*

Contents

ABSTRACT Determining an optimal training difficulty level for the best learning outcome is a crucial goal for adaptive educational systems. The literature supports the Inverted U-shape Hypothesis, suggesting that the ideal challenge level for learning is neither too easy nor too difficult. However, this optimal point depends on the type of training and response modality and may vary across domains, necessitating thorough examination before implementing adaptive learning procedures. This study aimed to investigate the influence of training difficulty on the learning outcomes of French medical students. Using data from a national educational platform, we explored the influence of the mean question difficulty encountered during training, relative to individual student ability, on the learning outcomes of medical students across diverse medical specialties. Importantly, the mean difficulty level varied randomly between students on this platform, mirroring a quasi-experimental design and enabling a thorough exploration of these effects. We first employed the Elo rating system to estimate the difficulty of platform questions and the evolution of students' abilities. A linear mixed-effects model was then used, with final exam performance as the main outcome and mean relative question difficulty during training (linear and quadratic terms) as the main predictor. Results showed a significant negative quadratic effect of mean relative difficulty on the final exam performance, revealing optimal difficulty levels for each medical specialty. Additionally, the analysis demonstrated that students with high abilities displayed a more pronounced inverted U-shaped relationship between training difficulty and final exam scores. This study advances our understanding of optimal training difficulty in the complex realm of medical education, by emphasizing the need to acknowledge variability across medical specialties and student abilities.

1 Introduction

The pursuit of knowledge, from initial acquisition to continuous practice, requires careful consideration of various factors influencing learning. Previous research has consistently highlighted the pivotal role of learners' retrieval practice behavior in predicting learning outcomes (Hattie, 2012; Henry L. Roediger III et al., 2006; T. Sinha et al., 2014; De Morais et al., 2014; Schütt et al., 2023). Thus, researchers and educators are keen on identifying conditions that promote optimal training and, consequently, optimal learning. Among multiple factors such as learner motivation, prior knowledge, feedback, learning schedule, learning tasks, etc., training difficulty— emerges as particularly crucial.

Searching for the ideal balance between challenge and achievability in training has been a central theme in educational research for many years. The Zone of Proximal Development idea of L. Vygotsky (1978) laid the foundation for this ongoing investigation. The theory proposed that the best learning occurs at the edge of the learner's competence, and research has shown that teaching skills within this zone lead to better performance compared to those outside it (Zou et al., 2019). This zone encourages learners to cope with new concepts while remaining achievable enough to prevent discouragement and frustration. Yet, given that each learner possesses different skills and levels of prior knowledge of the learning material, Vygotsky's theory inherently implies that the term "optimal difficulty" is a value that depends on the learner's ability. Thus, in practice, its original definition provides little guidance to determine an adequate difficulty level.

In an attempt to provide more concrete guidelines to achieve this balance and operationalize the concept of optimal difficulty, Rosenshine (2012) famously claimed that an 80% success rate in guided practice is optimal for learning. However, this claim was predominantly grounded in limited and small-scale evidence. It originated from observational "process-product" studies in elementary school reading and mathematics (Good et al., 1977; Gersten et al., 1982). Consequently, the reliability and generalizability of these conclusions across diverse educational contexts remain ambiguous and under-explored.

Due to the inherent variability in learners' competence, traditional approaches to determining the appropriate level of challenge in learning materials often relied on handcrafted rules devised by domain experts. However, the advent of Adaptive Learning Systems (ALS) (Park et al., 2003) has revolutionized this methodology by employing adaptive algorithms and extensive question banks to provide dynamic and personalized learning experiences. There are several examples of adaptive online learning systems e.g. for learning factual knowledge in the field of geography (Pelánek et al., 2017) or mathematics (Klinkenberg et al., 2011). Research consistently highlights the superiority of ALS over non-adaptive counterparts, demonstrating markedly improved learning outcomes (Papoušek, Stanislav, et al., 2016a; Ma et al., 2014).

In particular, adjustment of the difficulty level in training exercises to accommodate diverse prior knowledge and learning rates of learners holds significant promise. Previous research has suggested that selecting the appropriate difficulty level for practice exercises yields positive effects on how students perceive their own ability (Power, 2019), learning gains (Sampayo-Vargas

et al., 2013) and learning experiences (Kostons et al., 2010; Deterding et al., 2023). Consequently, various methods have been explored to adjust difficulty levels according to the user's needs. One approach involves granting users autonomy to select their next difficulty level—a process known as self-determined adjustment (Westlin et al., 2019). Alternatively, another method involves estimating the user's ability level based on their past performance and automatically presenting exercises tailored to that level. For example, presenting more challenging exercises to more advanced or high-performing students and simpler exercises to beginners or those who struggle. This process is often referred to as Dynamic Difficulty Adjustment (DDA). Research has underscored the benefits of both integrating learner choice into learning activities (Salden et al., 2006) and implementing dynamic adjustment strategies (Romero et al., 2006; Sampayo-Vargas et al., 2013; Y. Zhang et al., 2021), resulting in enhanced learning outcomes across diverse domains. However, a recent study (Schütt et al., 2023) comparing dynamic difficulty adjustment with self-determined difficulty levels in open-ended learning tasks found no significant differences in learning gains. Additionally, a separate study (Papoušek and Pelánek, 2017) suggested that instead of providing learners with the option to adjust difficulty, developing methods for automatic adjustment of target difficulty may offer a more effective solution.

Despite the evident advantages of adaptive learning systems that customize difficulty levels based on individual learner abilities, a fundamental question persists: what defines an appropriate level of difficulty to target? In many adaptive learning systems, these target difficulty levels are externally set by developers based on various considerations.

One example of this approach is the staircase method, which adjusts task difficulty to maintain a fixed error rate during learning, typically targeting an accuracy range of 80–85% (García-Pérez, 1998). This choice, however, is primarily based on intuition rather than empirical validation. Furthermore, enforcing a single target difficulty level applicable to all learners, contents and contexts may not be optimal. At the very least, the optimal level should depend on the chance level of correct responses, which is influenced by the question/response format. Additionally, it may be influenced by domain complexity, student skill, and other context-dependent factors. For instance, an 80% success rate may not represent the same level of difficulty across different learning activities, such as problem-solving tasks, generative activities with open-ended responses, and multiple-choice questions with varying numbers of propositions.

These complexities underscore the critical need for a more rigorous investigation into the validity and effectiveness of predetermined challenge levels. Therefore, before integrating an adaptive algorithm into learning systems, a thorough exploration of the optimal training difficulty becomes indispensable for maximizing learning outcomes in a given learning task and platform.

Recent studies exploring optimal training difficulty have provided support for the Inverted-U shape Hypothesis . A large-scale observational study (Abuhamdeh and Csikszentmihalyi, 2012) involved thousands of online chess players. The results revealed that players derived the greatest enjoyment when they competed against opponents who were slightly more skilled than they were (with a winning probability of approximately 20%), reflecting a correlation with the Inverted-U pattern. One notable limitation of this study was that players were not randomly

assigned opponents; instead, they had the autonomy to select opponents based on their chess rank, allowing them to control their expected game difficulty. Furthermore, the study focused on an intrinsically motivated activity, and the outcome of interest was the engagement of online chess players. It therefore says nothing about what they learned. In contrast, school and work-related activities often lack intrinsic motivation, being driven by an obligation to learn. Thus it is important to bear in mind that the optimal difficulty for enjoyment and motivation may not be the same as those for learning. Notably, Cao et al. (2022) demonstrated that motivation plays a mediating role in the relationship between difficulty and learning in educational gamification. Nonetheless, the optimal conditions for activities with intrinsic motivation may not be the same as those for less intrinsically motivated activities.

In educational contexts, evidence suggests that the impact of difficulty on learning and engagement can be conflicting. In the context of their simple educational game, D. Lomas et al. (2013) found that simpler problems resulted in higher engagement but lower levels of learning. Similarly, research conducted using the Math Garden software Jansen et al. (2013) compared three conditions with target success rates of 60%, 75%, and 90%, revealing that the easiest condition resulted in the most effective learning (mediated by the number of solved tasks). Another recent study (Papoušek, Stanislav, et al., 2016b) examining question difficulty in the learning of declarative knowledge arrived at a different conclusion, suggesting that more challenging questions enhance learning and long-term engagement, while easier questions are better for short-term engagement. Thus, while there is a general understanding that tasks are most engaging when they strike a balance between being too easy and too hard, the precise optimal point and the specific impact of a one-point increase in training difficulty on the learning outcome remain open questions.

Recently, the concept of optimal difficulty was extended beyond human learning in the domain of machine learning. R. C. Wilson et al. (2019) provided evidence that an 85% training accuracy served as a "sweet spot" for optimal learning in binary classification tasks. This theoretical result aligns with the inverted U-shape hypothesis, and supports the idea that relatively easy tasks promote better learning, consistent with Jansen et al. (2013) but not with D. Lomas et al. (2013) and Papoušek, Stanislav, et al. (2016b).

Collectively, these studies emphasize a key principle: optimal learning occurs when difficulty is calibrated appropriately. Nevertheless, diverse findings suggest that a universal optimal difficulty may not exist for all individuals and across every domain.

Therefore, this study aimed to provide a robust answer to the question of optimal training difficulty, within the specific context of multiple-choice questions in a complex domain: medical education. It relies on an online non-adaptive training platform that encompasses an extensive knowledge corpus, significant overlap between knowledge domains, and a very large bank of multiple-choice questions. We specifically addressed the following two research questions (RQs):

- **RQ1:** Can an optimal level of training difficulty be determined for medical students using our online platform?
- **RQ2:** If so, what is it, and does it vary between medical specialties?

To address our research questions, we applied the Elo rating system (Elo et al., 1978) to our dataset in order to estimate both the proficiency levels of students as they trained and the difficulty levels of questions within the platform. We then capitalized on the fact that on this platform, training questions are drawn at random. Therefore, students are exposed to questions of random difficulty. Over a training semester, by pure chance, some students will be exposed to more difficult questions than others, on average. Therefore, the mean difficulty level of the questions on which students trained can be considered as an instrumental variable in a quasi-experimental design. Thus, we conducted statistical analyses to examine the impact of the mean training difficulty (relative to students' competence) on the final exam performance, aiming to identify the optimal difficulty level that leads to the best learning outcome, as measured on a final exam.

We hypothesized that the relation between mean training difficulty and exam performance is non-linear, with a maximum at an optimal difficulty level. Moreover, recognizing the inherent diversity in difficulty levels across medical specialties, we further hypothesized that the optimal difficulty level might differ to some extent between specialties. Of course, we also expect the main effects of student ability and training intensity on exam performance. Since these factors may affect both training and final performance, they will be included in our analyses as covariates.

2 Methods

2.1 Study Design and Platform Features

This study adopted a quasi-experimental research design, utilizing data sourced from the Banque Nationale d'Entraînement (BNE) digital learning platform. The subsequent section provides a comprehensive description of the BNE platform, its operational mechanisms, and the specific features of the question bank. A detailed presentation of the dataset is given in the following section.

2.1.1 BNE Digital Learning Platform

The Banque Nationale d'Entrainement (BNE) digital learning system is an online platform managed by Université Numérique en Santé et en Sport (uness.fr) that is widely used by more than 8000 medical students from all French universities in each academic year from the 2nd to the 6th year. It is used for the administration of exams as well as for student training. The training on this platform is entirely student-driven and involves tackling multiple-choice questions across 31 different medical specialties, which helps students test and reinforce their knowledge in specific areas of medicine and prepare for the ECN (Épreuves Classantes Nationales) national final exam. This high-stakes exam, which takes place at the end of the sixth academic year, is the main determinant of students' access to residency in their chosen specialty.

The platform also incorporates a feedback system designed to enhance the learning experience for students. After each test, students receive corrective feedback on the correctness of their answers and, in a subset of questions, detailed explanations of why some answers are right or

wrong.

Concretely, students who initiate a training session on the BNE digital learning platform freely select 1) the type of questions (isolated questions, progressive medical files, critical article reading, see description below); 2) the medical specialties they want to train on; 3) the number of questions of each type in each specialty. They can generate multiple training sets and train at their own pace as long as they want. This flexibility enables them to tailor their training program to their individual preferences and needs.

Importantly, although students are free to select the desired specialties and test types for training, the specific questions of each type and specialty are drawn at random from the question bank. Thus, students may be exposed by chance to questions of higher or lower difficulty on average. The mean difficulty level of the questions on which students trained can therefore be considered as an instrumental variable in a quasi-experimental design. We exploit this random allocation to test the effect of mean question difficulty during training (relative to students' current ability) on final exam outcomes.

2.1.2 Question Bank

Past exam questions, along with those designed for training purposes, are incorporated into a comprehensive question bank, constituting an extensive collection of approximately $2,313,023$ multiple-choice questions across 31 medical specialties. This huge question bank is easily accessible to students within the platform, offering a valuable and rich training resource to prepare for the ECN. The training tests drawn from the question bank on the platform are categorized into three main types:

1. **Isolated Questions Test (IQ):** Classic multiple-choice questions.
2. **Progressive File Test (PF):** A set of approximately 15 interconnected multiple-choice questions, where each question reveals new medical information about a patient, thus simulating a realistic medical situation.
3. **Critical Article Reading Test (CAR):** A set of questions based on an article.

Furthermore, within each of these test categories, questions may be either SAQs (Single Answer Questions) or MAQs (Multiple Answer Questions).

It is also worth highlighting that the average number of specialties linked to each question is 1.58. This underscores that a considerable portion of questions (34.5%) in the question bank are tagged with multiple medical specialties, implying that answering them accurately requires knowledge from several medical specialties. Another notable aspect of the question bank is that, while it comprises questions with varying numbers of propositions, a significant majority of the questions — 98.2% of training questions and 97.3% of ECN questions — consist of five propositions.

2.2 Dataset

We extracted detailed records of student interactions from the BNE, including every attempt made by students to answer questions. The data included anonymized student IDs, question IDs, specialties tagged by the question, timestamp of the attempt, number of propositions in the question, test type, and score obtained in each attempt. This approach ensures the privacy and confidentiality of all students involved.

To test our hypothesis, we focused our analysis on the educational year 2020-2021, which was the most recent accessible dataset at the time of our analysis. Although direct access to the official ECN national final exam via the BNE platform was not available, we had access to a final exam known as ECNp (ECN préparatoire). This mock exam, typically held in mid-March (specifically on March 15th, 16th, and 17th, 2021, for the 2020-2021 academic year), closely mirrors the format of the actual ECN national final exam, and therefore served as an external outcome identical for all 6th year students. We therefore focused on 6th-year students, and we will now refer to this mock exam as the final exam.

In order to obtain a dataset allowing for a robust estimation of question difficulty and student ability in each specialty, and allowing for a robust analysis of the relation between training conditions and exam results, we applied a number of steps of data cleaning and selection, as described below.

2.2.1 Data Cleaning

After obtaining the raw data for the 6th-year students for the 2020-2021 educational year, we first conducted a series of preprocessing steps on this dataset. These steps were executed in the following sequence:

1. Removal of duplicated rows: Only 144 duplicated rows out of 27,557,031 were removed.
2. Exclusion of questions without any tagged medical specialty: 0.28% of the questions (making 0.21% of attempts) lacked specialty tags.
3. Binarization of question grading: BNE employs a sophisticated grading scheme depending on the number of correct and incorrect answers ticked. We binarized scores by mapping all scores other than 1 to 0.

2.2.2 Training Period Dataset

Table 3.1 provides an overview of the key characteristics of our training period dataset.

An essential variable highlighted in Table 3.1 is the student-question sparsity, indicating the proportion of missing values in the student-question interaction matrix. The table illustrates that the dataset is extensive but shows significant sparsity (*Sparsity (student, question)* = 0.99) during the training period. This reflects that there are considerably more available questions than any individual student can attempt, and different students most often attempt different randomly selected questions.

2.2.3 ECNp - Final Exam Dataset

Table 3.1: Data Set Summary

Data	Period	Total Nb Attempts	Nb Students	Nb Questions	Nb Specialties	Sparsity (Student, Question)
Training Period Dataset	16.11.2020–14.03.2021	18,503,110	8,633	323,701	31	0.99
Final Exam (ECNp) Dataset	15–17.03.2021	3,178,527	8,636	372	28	0.01

Table 3.1 also shows characteristics of the final exam (ECNp) covering 28 different medical specialties (addictology, orthopedics, and toxicology were not tested in the final exam). A total of 3,178,527 user interactions were recorded, with a sparsity value of 0.01, suggesting that nearly all students attempted almost every question in the exam.

It is important to emphasize that the final exam is not a retest of training questions, but rather a brand new set of questions designed to test students' ability to generalize their medical knowledge acquired in training to unfamiliar questions. This reflects the actual ECN exam's focus on broader understanding over rote memorization. This differs from many other experimental studies where learning is assessed mainly on repeated questions.

It is also noteworthy that the distribution of questions across medical specialties in the final exam is not uniform. Figure 3.1 illustrates the varying number of questions allocated to each specialty in the final exam, along with the corresponding proportion of correct answers.

2.2.4 Data-filtering for Statistical Analysis

Although the Elo rating system was initially applied to unfiltered data to obtain the mean relative difficulty during the 4-month training period for the total of the 253,036 student-specialty pairs, for the statistical analysis we implemented filtering procedures to enhance the reliability of results.

Firstly, to ensure a robust measure of learning outcome (final exam score), the analysis was limited to medical specialties with a sufficient number of questions ($\geq$ 20) in the final exam, resulting in 13 out of 31 specialties being included (see Figure 3.1).

Note that all students did not necessarily train on all specialties and did not always take enough questions in each specialty to allow for a reliable estimation of their ability in all specialties. Therefore, each student's training can only be analysed on a specific subset of medical specialties. Student-specialty pairs are thus the unit of our statistical analysis. Since our analysis focuses on within-subject differences between medical specialties, we excluded students who had not undergone training in at least two of the 13 selected medical specialties, with a minimum of 100 attempts each, to ensure a reliable estimation of online student ability from Elo rating system. This led to the removal of 15% of students from the analysis.

As a result of these multiple data cleaning and selection steps, the final dataset available for analysis included 6,451 students, resulting in 45,436 student-specialty pairs and a 46% sparsity of the student-specialty matrix.

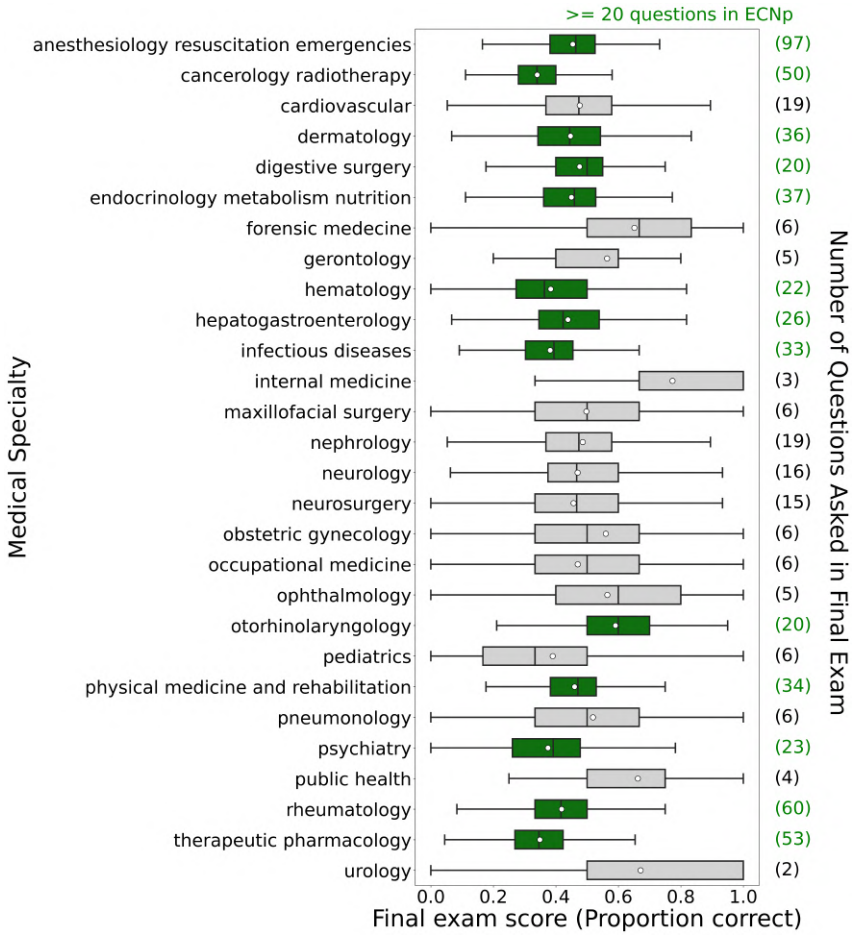


Figure 3.1: Horizontal box plot illustrates the variability in correct answer rates among different medical specialties within the ECNp final exam. The parentheses on the right y-axis denote the number of questions posed in each specialty within the ECNp exam. Only specialties with a sufficient number of questions (≥ 20), shown in green, are included in the analysis.

2.3 Students' Ability and Question Difficulty Estimations

Traditionally, models for students' ability and question difficulty estimations trace back to Rasch's single-parameter logistic regression formulation (Rasch, 1960). Item Response Theory (IRT), based on this logistic model and its variations (Linden et al., 2013), has been extensively used in educational applications, providing robust estimations (for a recent overview, see Wauters, Desmet, and Van den Noortgate (2010)).

While IRT is a well-established method, alternative approaches for estimating task difficulty and student ability have emerged. These include Markov process models (Gweon et al., 2015), deep knowledge tracing models (Chan et al., 2021; Piech et al., 2015), and rating systems (Pelánek, 2016), each documented in existing literature.

In our context, because we were analyzing training data over several months, it was important to have an up-to-date estimate of each student's ability each time they answered a new question, hence the need for dynamic estimations. This led us to focus on rating systems, and more particularly on the Elo rating system, as described by Pelánek (2016). The Elo rating system stands out as an algorithmically and computationally more economical approach compared to other models that necessitate processing vast amounts of previously collected data. Moreover, it has found widespread application in educational technologies (Klinkenberg et al., 2011; Pelánek, 2016; Attali, 2014) and has been validated as a suitable online tool without the need for extensive question pretesting on large sample sizes (Papoušek, Pelánek, and Stanislav, 2014; Antal, 2013; Kandemir et al., 2024).

2.3.1 Elo Rating System for Question Difficulty and Online Student Ability Estimations

Initially designed for ranking chess players, the Elo rating system (Elo et al., 1978) has found a new application in educational settings. In this context, students and learning materials (e.g. questions) are treated as opponents, each assigned ratings that reflect their abilities and difficulties, respectively. This repurposing allows the system to predict ratings for students and questions, functioning as an assessment tool for both student ability and question difficulty in educational environments.

In its standard educational formulation, every student s has an ability parameter θ_s and each question q has a difficulty parameter θ_q. The probability of student s successfully answering a multiple-choice question q, denoted as $\Pr(a_{sq} = 1|\theta_s, \theta_q)$, can be obtained using a logistic function σ of the discrepancy between student ability and question difficulty:

$$\Pr(a_{sq} = 1|\theta_s, \theta_q) = \sigma(\theta_s - \theta_q) = \frac{1}{1 + e^{-(\theta_s - \theta_q)}} \tag{3.1}$$

The system then dynamically adjusts the student's ability and the question's difficulty parameters after each interaction based on the disparity between the estimated probability correct ($Pr(a_{sq} = 1)$) and the actual score (a_{sq}).

These updates are defined by the following formulas for question difficulty (θ_q) and for

student ability (θ_s):

$$\begin{aligned} \theta_q &:= \theta_q + K(\Pr(a_{sq} = 1|\theta_s, \theta_q) - a_{sq}) \\ \theta_s &:= \theta_s + K(a_{sq} - \Pr(a_{sq} = 1|\theta_s, \theta_q)) \end{aligned} \tag{3.2}$$

where K is a constant value that determines the degree of update sensitivity based on the student's most recent attempt.

Thus the Elo rating system continuously adjusts student and question parameters through ongoing interactions. This system provides both current estimations of difficulty and ability parameters after each interaction, as well as final versions of these parameters at the end of the simulation. In this paper, we refer to the dynamically estimated ratings after each attempt as online difficulty and online student ability, and to the ratings stabilized at the end of the training as final student ability and final question difficulty.

Following the establishment of the standard Elo rating system formulation, the system has undergone various extensions (Pelánek, 2016; Abdi, Khosravi, and Sadiq, 2021). In a prior investigation, the multivariate version of the Elo rating system was adapted to suit the current database (Kandemir et al., 2024). This version accommodates situations where learning items can be tagged with multiple knowledge components, enabling a more nuanced estimation of student abilities. Similar to the standard Elo rating system, the predictive efficacy of the multivariate version has been established in previous studies (Abdi, Khosravi, and Sadiq, 2021; Kandemir et al., 2024). These studies, drawing upon data from varied learning platforms, underscore its reliability in accurately estimating both student and question-related parameters, thereby highlighting its robustness.

As mentioned before, in the BNE, one question may be tagged by multiple specialties. Thus, to effectively account for student abilities within each of these tagged specialties, we employed the multi-concept extended version of the Elo rating system, as introduced by Abdi, Khosravi, and Sadiq (2021). The estimation of a student's online ability on a question involved averaging the student's online abilities across all specialties linked to the question. The details and exact formulation of the multivariate Elo rating system adapted to our dataset have been described in a previous work (Kandemir et al., 2024). Therefore, in this current study, we will utilize it as a "black box" model, focusing on its applications and outputs rather than its internal workings. The predictive efficacy of this multivariate Elo rating system within our dataset has also been evaluated in the aforementioned previous study. This was achieved by training the model on training period data and subsequently utilizing it to predict the outcomes of mock final examinations. The high predictive performance (AUC = 0.81, RMSE = 0.42) aligns with findings from previous research (Abdi, Khosravi, and Sadiq, 2021; Klinkenberg et al., 2011; Wauters, Desmet, and Van Den Noortgate, 2011). Here, the Area Under the Curve (AUC) measures the model's ability to distinguish between classes, while the Root Mean Square Error (RMSE) reflects the average magnitude of prediction errors, with smaller values indicating better predictions.

2.4 Measures

Following the derivation of estimations for both students' abilities and question difficulties using the Elo rating system on the training data, we computed the following measures for each student-specialty pair:

Mean Relative Difficulty (Main predictor): In order to measure how challenging a student's training period was, we calculated the relative difficulty of each attempt by subtracting the student's estimated online ability on the attempt (averaged across all specialties tagging this particular question) from the estimated final difficulty of the question in the training period dataset. Whereas it is important to take online student ability at each attempt into account because it is expected to improve with practice, we used final rather than online question difficulty because it is presumed to be a stable property whose estimation is most reliable after many attempts. We then averaged the relative difficulty for each attempt that the student made on questions from a given specialty over the entire training period, yielding a mean relative difficulty for each student-specialty pair.

For a student S who attempted n questions in a given specialty k:

$$\text{Mean Relative Difficulty}_{S,k} = \frac{1}{n}\sum_{i=1}^{n}(\delta_i - \theta_{s,k,i})$$

Where:

- δ_i: The final difficulty estimation of question i, considered stable due to many attempts
- $\theta_{s,k,i}$: The online ability of student S in specialty k, which is required to solve question i, at the time of the attempt

For example, if a student with an online ability of 1 (positive values representing higher ability) took a question with a difficulty rating of 2 (positive values representing more difficult questionss), the relative difficulty of this action would be $(2 - 1 = 1)$. A positive mean relative difficulty score indicates that a student was exposed to relatively challenging questions during their training period for the given specialty, while a negative score suggests a relatively easier training period.

Final Exam Score (Main outcome): The learning outcome was the proportion of correct responses obtained by each student in each specialty in the final exam. It is calculated by using the Final Exam (ECNp) Dataset.

Given the observational nature of the study design and the reliance on Elo rating system estimates for both student ability and question difficulty, controlling for all potential dependencies —such as students' interactions with the platform and their inherent abilities—presented a challenge. One notable concern was the potential correlation between the mean relative difficulty and the final student ability. It was anticipated that higher-ability students would, on average, encounter lower mean relative difficulty during training. Consequently, a positive association between mean relative difficulty and exam score was likely, as higher-ability students tend to achieve higher scores in the final exam (Hattie, 2012). In order to control for this confounding factor, our statistical model was adjusted on final student ability in each specialty. Additionally,

we accounted for training intensity, measured by the number of questions attempted per specialty during training, as prior research has demonstrated that student engagement with practice questions significantly impacts learning outcomes (Lin et al., 2019).

Final Student Ability (Covariate): The final student ability in each medical specialty was the last estimation of their ability at the end of the training period. It is calculated by using the training period dataset.

Number of questions taken by specialty (Covariate): To account for the impact of training intensity on final exam outcomes, we included the number of questions taken by each student in each specialty and its interaction with final student ability as covariates in our model. This allowed us to better take into account the combined effects of training intensity and final student ability on exam performance. It is calculated by using the training period dataset.

After calculating these measures, we combined them for each student-specialty pair. Subsequently, a data-filtering process (refer to Section 2.2.4) was applied to refine this combined dataset. The resulting filtered data was then used for the statistical analysis.

2.5 Analyses

2.5.1 Statistical Model

We employed a linear mixed-effects model to investigate the influence of mean relative question difficulty on final exam scores. This analysis was conducted using the `lme4` package (Bates et al., 2015) (Version 1.1-31) in R (Version 4.2.2), and the restricted maximum likelihood (REML) estimation method to account for the hierarchical nature of the data. Additional packages used for data wrangling, visualization, and diagnostics included `dplyr` (Wickham et al., 2023), `ggplot2` (Wickham, 2016), `performance` (Lüdecke et al., 2021), `sjPlot` (Lüdecke, 2021), and `lmerTest` (Kuznetsova et al., 2020). A complete list of all packages used, along with their versions, is provided in the openly shared repository.

The Final Exam Score was regressed on both linear and quadratic Mean relative difficulty as fixed effects, to test the hypothesis of a non-linear effect with a maximum.

Given the observational nature of the study design and the reliance on Elo rating system estimates for both student ability and question difficulty, controlling for all potential dependencies posed a challenge. One notable concern was the potential correlation between mean relative difficulty and student ability. Since relative difficulty is the difference between question difficulty and student online ability, it was anticipated that students with higher ability would, on average, end up with lower mean relative difficulty during training. Consequently, a negative association between mean relative difficulty and exam scores was expected, as students with higher final ability would tend to achieve higher scores in the final exam. In order to control for this confounding factor, our statistical model was adjusted on final student ability in each specialty.

An interaction between final student ability and both the linear and quadratic terms of mean relative difficulty was included to explore whether the impact of mean relative difficulty varied with final student ability.

The model also included the number of questions taken by each student in each medical specialty as a covariate. Moreover, we included an interaction between final student ability and the number of questions taken by students in each medical specialty, in order to test the hypothesis that training intensity may have a different effect depending on final student ability. To facilitate the interpretation of interaction terms, the number of questions taken in each specialty was mean-centered, while other predictors in the model were already centered around a mean of approximately zero (see Table 3.2).

We addressed the within-student between-specialty design and the complex structure of our dataset by further including:

Random Intercepts for Students. Capturing inherent variability in student performance across the dataset, independent of the specialty-specific influences.

Random Intercepts and Slopes for linear and quadratic terms of Mean Relative Difficulty. This allowed the influence of mean question difficulty on exam scores to vary across specialties.

Our linear mixed-effects model was thus formulated as follows (in R syntax):

$$\begin{aligned} \text{final\_exam\_score} \sim & \\ & \text{mean\_relative\_difficulty} * \text{final\_student\_ability} + \\ & \text{mean\_relative\_difficulty}^2 * \text{final\_student\_ability} + \\ & \text{nb\_questions\_taken\_by\_specialty} * \text{final\_student\_ability} + \\ & (1|\text{student}) + \\ & (1 + \text{mean\_relative\_difficulty} + \text{mean\_relative\_difficulty}^2|\text{specialty}) \end{aligned}$$

3 Results

3.1 Descriptive statistics

Table 3.2 displays the descriptive statistics of the variables under analysis.

We assessed potential multicollinearity among all predictor variables using the Variance Inflation Factor (VIF). A VIF value exceeding the commonly accepted threshold of 5 indicates potential multicollinearity issues (Kim, 2019; Sheather, 2009). The results, presented in Table 3.3, show that all predictor variables included in our model have VIF values between 1.17 and 3.83. These values suggest that our model does not suffer from significant multicollinearity concerns.

3.2 Effects of Training Difficulty on Learning Outcome

Results of the Linear Mixed-Effects Model Analysis are provided in Table 3.4, with correlations between fixed effects in Table 3.5. The main effect of mean relative difficulty (Estimate: -0.05, 95% CI: [-0.06, -0.04], $p < 0.001$) indicates that, when final student ability is held constant at

Table 3.2: Descriptive statistics

Filtered data used in the model:	
Number of students	6,451
Number of medical specialties	13
Number of student-specialty pairs	45,436
Sparsity in the student-specialty matrix	46%
Descriptive statistics of the model variables:	
	M **SD**
1. Mean relative difficulty	-0.09 0.43
2. Final student ability	0.07 0.41
3. Number of questions taken by specialty	280 236
4. Final exam score (proportion correct)	0.45 0.13

Table 3.3: Multicollinearity Check Results

Term	**VIF**	**VIF 95% CI**	**Increased SE**	**Tolerance**	**Tolerance 95% CI**
Mean relative difficulty	1.27	[1.25, 1.28]	1.13	0.79	[0.78, 0.80]
Number of questions taken by specialty	1.32	[1.31, 1.34]	1.15	0.76	[0.75, 0.76]
Mean relative difficulty^2	3.80	[3.74, 3.86]	1.95	0.26	[0.26, 0.27]
Final student ability	1.23	[1.22, 1.24]	1.11	0.81	[0.80, 0.82]
Mean relative difficulty : Final student ability	3.83	[3.77, 3.89]	1.96	0.26	[0.26, 0.27]
Mean relative difficulty^2 : Final student ability	1.17	[1.15, 1.18]	1.08	0.86	[0.85, 0.87]
Number of questions taken by specialty : Final student ability	1.37	[1.35, 1.39]	1.17	0.73	[0.72, 0.74]

its mean (0.07), a 1 standard deviation increase in mean relative difficulty is associated with a 5 percentage point decrease in final exam scores. Furthermore, the final exam score was negatively associated with the squared mean relative difficulty (Estimate: -0.03, 95% CI: [-0.04, -0.02], $p < 0.001$), indicating that the relationship between mean relative difficulty and exam scores was curvilinear with a maximum (see Figure 3.2). This supports the existence of an optimal difficulty level at which exam scores reach a maximum.

Table 3.4: Linear Mixed-Effects Model Analysis Results: Fixed Effects

Fixed Effects	**Estimates**	**Std. Error**	**95% CI**	***p*-values**
(Intercept)	0.46	0.02	[0.42, 0.50]	4.72×10^{-11} ***
Mean relative difficulty	-0.05	0.01	$[-0.06, -0.04]$	7.50×10^{-8} ***
Mean relative difficulty^2	-0.03	0.01	$[-0.04, -0.02]$	4.30×10^{-6} ***
Final student ability	0.05	0.00	[0.04, 0.05]	2.51×10^{-70} ***
Number of questions taken by specialty	2.53×10^{-5}	2.64×10^{-6}	$[2.01 \times 10^{-5}, 3.05 \times 10^{-5}]$	1.09×10^{-21} ***
Mean relative difficulty : Final student ability	-0.02	0.01	$[-0.03, -0.01]$	2.68×10^{-4} ***
Mean relative difficulty^2 : Final student ability	-0.01	0.00	$[-0.01, -0.00]$	6.43×10^{-5} ***
Number of questions taken by specialty: Final student ability	-7.00×10^{-6}	4.84×10^{-6}	$[-1.65 \times 10^{-5}, 2.48 \times 10^{-6}]$	1.48×10^{-1}
N specialties	13			
N students	6451			
N observations	45436			

Significance Codes: '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$.

Table 3.5: Linear Mixed-Effects Model Analysis Results: Correlation of Fixed Effects

Correlation of Fixed Effects	1	2	3	4	5	6	7	8
1. Intercept	1.00							
2. Mean relative difficulty^2	-0.347	1.00						
3. Mean relative difficulty	-0.406	0.145	1.00					
4. Number of questions taken by specialty	0.011	0.013	0.064	1.00				
5. Final student ability	0.002	0.033	0.361	-0.031	1.00			
6. Mean relative difficulty^2 : Final student ability	0.006	-0.003	0.105	0.040	-0.157	1.00		
7. Mean relative difficulty : Final student ability	0.000	0.835	0.052	0.000	0.037	0.137	1.00	
8. Number of questions taken by specialty : Final student ability	-0.004	-0.020	-0.040	-0.487	0.059	-0.050	0.077	1.00

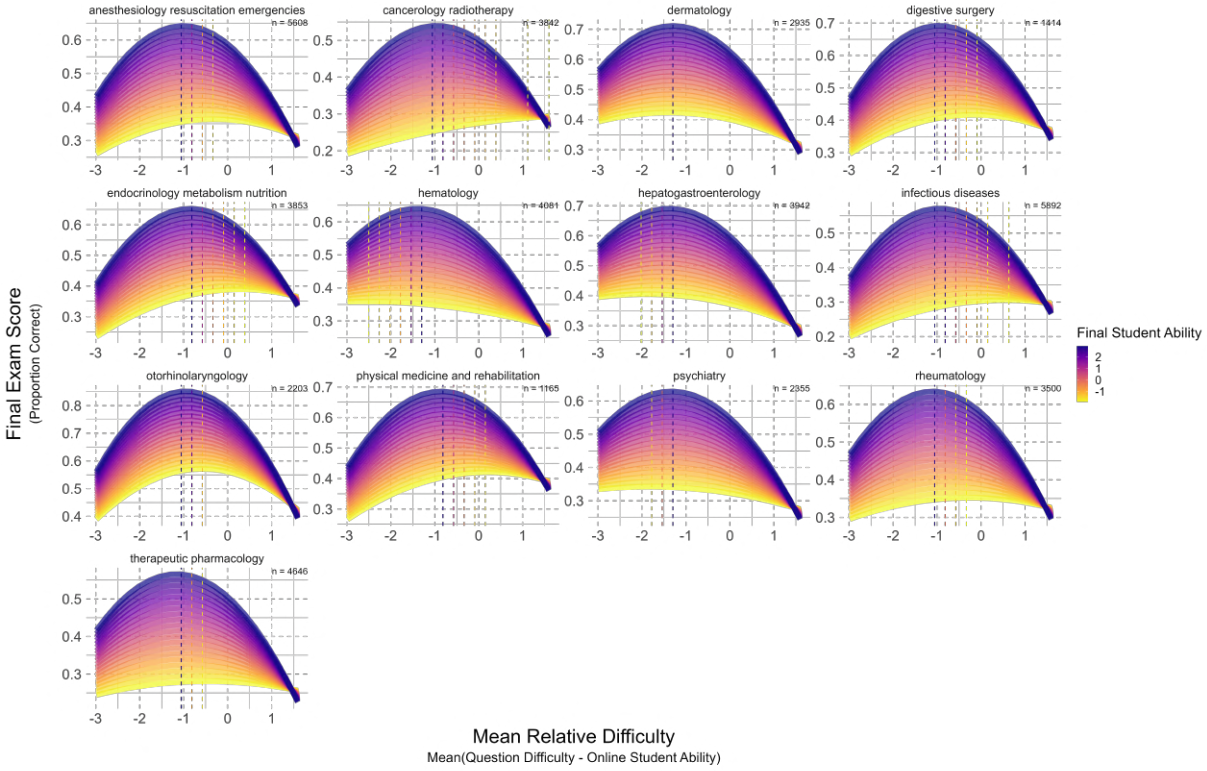


Figure 3.2: Model fit by specialty: Relation between Mean Relative Difficulty and Final Exam Score (proportion correct) across levels of Final Student Ability, with the Number of questions taken by specialty fixed at the mean value of 280. The color gradient represents the spectrum of final student abilities, ranging from low (yellow) to high (blue). Vertical dashed lines within each panel, which use the same color gradient as the curves, indicate the optimal mean relative difficulty (e.g., maxima of the curves) for different levels of final student ability. Additionally, the sample sizes ('n') refer to the number of students included in the model fit for each specialty.

3.3 Differential Effects of Training Difficulty on Learning Outcomes Across Student Abilities

Both final student ability (Estimate: 0.05, 95% CI: [0.04, 0.05], $p < 0.001$) and number of questions taken by medical specialty (Estimate: 2.5×10^{-5}, 95% CI: $[2.0 \times 10^{-5}, 3.0 \times 10^{-5}]$, $p < 0.001$) positively and significantly affect final exam scores, with their effects reflecting relationships with exam performance when the mean relative difficulty is held constant at its average value (-0.09). This shows that controlling for mean relative difficulty during training, students with higher ability and those who train harder tend to achieve superior scores, as expected.

Figure 3.2 illustrates the model fit, showing the effects of mean relative difficulty (x-axis) and final student ability (color-coded) on the final exam score (y-axis) across different specialties.

Moving beyond the main effect of final student ability, we found a statistically significant interaction between final student ability and mean relative difficulty (Estimate: -0.02, 95% CI:[-0.03, -0.01], $p < 0.001$). This suggests that the optimal difficulty level significantly varied depending on students' ability as illustrated in Figure 3.2 by the shifting pattern of vertical lines, color-coded to represent different ability levels within the same specialty. Although this pattern is not uniform across all specialties, higher-ability students (dark blue) generally have a relatively lower optimal relative difficulty level. Furthermore, there also was an interaction between the squared mean relative difficulty and final student ability (Estimate: -0.01, 95% CI:[-0.01, -0.00], $p < 0.001$), indicating that high-ability students exhibited a greater curvature of the inverted U-shaped relationship between mean question difficulty and final exam scores (as shown by the color gradient of the curves in Figure 3.2), and therefore showed greater sensitivity to the mean relative difficulty of questions they were exposed to.

3.4 Optimal Training Difficulty Differs between Medical Specialties

Figure 3.2 demonstrates that the apex of the curves, representing the optimal relative difficulty level for achieving the highest final exam score, varies across specialties. To test the significance of the specialty effect on the apex of the curve, we examined the significance of the random slopes for mean relative difficulty within the specialty grouping factor in our mixed-effects model (Table 3.6). This was achieved by conducting a likelihood ratio test by using the maximum likelihood (ML) estimation method to compare the full model with a reduced model in which the random slope of mean relative difficulty within the specialty grouping factor was excluded. The comparison, shown in Table 3.7, reveals a significant improvement in fit upon the inclusion of the random slope of mean relative difficulty within the specialty grouping factor in the Full Model ($\Delta\chi^2 = 339.64$, $\Delta df = 3$, $p < 0.0001$). Additionally, both AIC and BIC are lower for the Full Model, confirming the superior fit. This finding implies that the influence of mean relative difficulty on final exam scores differs between specialties.

As observed in Figure 3.2, there is a clear trend towards negative mean relative difficulties giving the maximum final exam scores. This pattern is consistent across different medical specialties and varies depending on the final abilities of the students. When adjusted for average

final student ability and the number of training questions, the optimal mean relative difficulties across specialties ranged from -1.55 to -0.53, with a mean of -0.952 (SD = 0.378). This suggests that, in this question bank, the best learning outcomes occur when students encounter, on average, slightly easier questions relative to their own ability.

Table 3.6: Linear Mixed-Effects Model Analysis Results: Random Effects

Random Effects		**Variance**	**Std. Dev.**	**Corr**	
Student	(Intercept)	0.00	0.06		
Specialty	(Intercept)	0.01	0.08		
	Mean relative difficulty	0.00	0.02	-0.47	
	Mean relative difficulty^2	0.00	0.01	-0.88	0.30
Residual		0.01	0.07		

Table 3.7: Comparison of Linear Mixed Models

Model	npar	AIC	BIC	logLik	Deviance	$\Delta\chi^2$	Δdf	p-values
Reduced Model	13	-96307	-96194	48167	-96333	-	-	-
Full Model	16	-96641	-96501	48336	-96673	339.64	3	$< 0.0001^{***}$

Note: p-values indicate the significance of the improvement in fit.
Significance Codes: '***' $p < 0.001$, '**' $p < 0.01$, '*' $p < 0.05$.

For easier interpretation of our findings, Figure 3.3 illustrates the correspondence between relative difficulty (x-axis) and probability of responding correctly (y-axis), using a logistic function transformation of relative difficulty, for various question types (single-answer vs. multiple-answer questions) and numbers of response propositions (ranging from 2 to 7). The curves shows that success probability decreases with relative difficulty following a logistic function. The comparison of the different curves reflects the obvious fact that single-answer questions are easier than multiple-answer questions (which have more possible responses, hence a lower chance level), and that questions with more propositions are more difficult than those with fewer propositions. For multiple-choice questions (MCQs) with five propositions—the most common question type in our question bank—the optimal relative difficulty range, adjusted for average final student ability and the average number of training questions across specialties, falls between -0.53 and -1.55 (indicated by red vertical lines). This range corresponds to a success probability of 0.64 to 0.83 (shaded area).

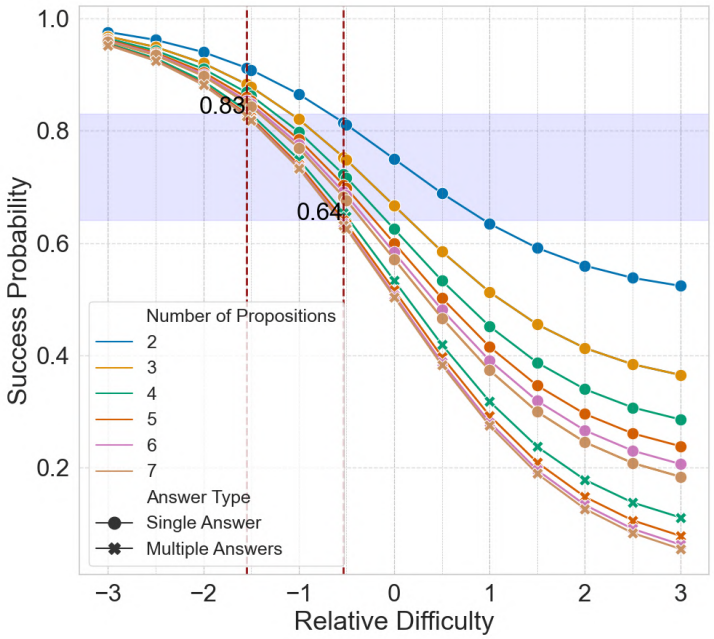


Figure 3.3: Correspondence between relative difficulty and success probability for questions of different types and number of propositions. The shaded area represents the optimal success probability range corresponding to a relative difficulty range of -0.53 and -1.55 for the most frequent 5-proposition multiple-answer questions with numerical values marking the boundaries.

4 Discussion

In this study focused on medical learning, we investigated the impact of training difficulty on learning outcomes.

To address our main research question (RQ1), we exploited the random allocation of training questions to medical students to use question difficulty as an instrumental variable. For each student and specialty, we computed the mean difference between question difficulty and online student ability during training (mean relative difficulty) and examined its influence on learning outcomes. Controlling for students' final ability in the medical specialty and the number of questions taken by specialty, our analysis reveals a significant quadratic relationship between mean relative difficulty and exam score. This finding supports the inverted U-shaped hypothesis, indicating the existence of an optimal difficulty level in this particular medical training setting. In our investigation of the second research question (RQ2), we analyzed the optimal difficulty level and its variation across different medical specialties and final student abilities. We found that the optimal difficulty level for achieving the best learning outcomes depends on both the final student's ability and the medical specialty in question.

In order to better appreciate the size of this inverted U-shape effect, one might consider a student of average ability (0.07) taking questions at his/her level, thereby experiencing a mean relative question difficulty of 0, in one particular specialty, say, anesthesiology (upper left in Figure 3.2). This corresponds to being correct about 50% of the time in multiple-answer questions (as shown in Figure 3.3). Following this training, this student would have

an expected score of about 45% at the final anesthesiology exam (from Figure 3.2). By changing his training regime to relatively easier questions (a relative difficulty of -1, corresponding to answering correctly 70% of the time), his expected score would improve to 47% (following the orange line on Figure 3.2). However, due to the interaction between student and linear and quadratic mean difficulty, a student of higher ability (maximum student ability=2.93, the dark blue line in Figure 3.2) would experience a greater benefit of shifting his training questions from a relative difficulty of 0 to -1: his predicted score would rise from 59% to 64%. Whereas a student at a lower ability level (minimum student ability=-1.92, yellow line) would actually see his predicted score decrease slightly by training on even easier questions. Finally, this pattern differs depending on specialty, as one can see in Figure 3.2 and as confirmed by the significance of the specialty random slope. These effects of a substantial change in training question difficulty on predicted exam scores may seem relatively small. However, it should be recalled that the specific characteristics of our dataset could only lead to minimizing effect sizes. Indeed, students trained both inside and outside the platform, but we had no information on the latter. Our analysis was restricted to the influence of training that was observable on the platform, potentially missing most of the training for some students. Furthermore, we didn't have any experimental control over training question difficulty. Thus students experienced very variable question difficulties, and the effect of mean question difficulty was limited by randomly occurring variation between students. Finally, outcomes were measured across broad medical specialties, assessing the far transfer of knowledge from training to the ECN exam, rather than rote memorization. Other educational settings where all the training occurs on the platform, where question difficulty is entirely controlled, and where the knowledge domains are narrower should expect larger effect sizes.

Our primary finding of an optimal difficulty level aligns with the intuitive understanding of many fundamental pedagogical principles, suggesting that effective learning occurs when tasks strike a balance between being neither too easy nor too difficult. Such an intuition has been formulated in prominent educational theories such as "zone of proximal development" by L. Vygotsky (1978), and Principles of instruction by Rosenshine (2012). Although quantification of success depends on task structure and chance level, our optimal success probability range of 64-83% is compatible with the 85% rule for optimal learning in stochastic gradient-descent based learning algorithms identified in the simulation work of R. C. Wilson et al. (2019). This optimal success probability of 64-83% (for 5-propositions multiple choice questions) that we observed may suggest that students learned best from very easy questions. This may actually reveal that the questions in the BNE medical training system are excessively difficult on average, and that adding easier questions, or adaptively adjusting question difficulty to final student ability would significantly benefit students. Implementing such changes would provide a more personalized and effective learning experience, consistent with prior research that highlights the benefits of tailoring training content (Zou et al., 2019; Maghsudi et al., 2021; Benvenuti et al., 2023).

Our results also conflict with some previous research. For instance, Jansen et al. (2013), in their study on the impact of training difficulty on math performance in children aged 8 to 13 years, found that the easiest difficulty level led to the most effective learning. They used

Table 3.8: Optimal relative difficulty by medical specialty for the most frequent 5-proposition multiple-answer questions

Medical Specialty	**Optimal Relative Difficulty**
Anesthesiology, Resuscitation, Emergencies	-0.8099
Cancerology, Radiotherapy	-0.5785
Dermatology	-1.3190
Digestive Surgery	-0.7174
Endocrinology, Metabolism, Nutrition	-0.5323
Hematology	-1.5504
Hepatogastroenterology	-1.5042
Infectious Diseases	-0.6248
Otorhinolaryngology	-0.8099
Physical Medicine and Rehabilitation	-0.5323
Psychiatry	-1.4116
Rheumatology	-0.9488
Therapeutic Pharmacology	-1.0413

a computer-adaptive program to adjust the difficulty of math problems to individual abilities, targeting success rates of 60%, 75%, and 90% for difficult, medium, and easy conditions, respectively. Furthermore, they identified a mediating effect of the number of problems attempted. Specifically, children in easier conditions attempted more problems in their platform, which indirectly led to greater math improvement. This mediation effect could explain the linear relationship between success rate and final performance observed in their study while our quasi-experimental study revealed a more curvilinear relationship. Additionally, the nature of the tasks in their study differed from ours. Their tasks required a single correct answer, which meant that success rates were straightforward to measure and control. In contrast, our study involved questions with multiple possible answers, adding complexity to the success rate measurement. This difference in task structure may contribute to the contrasting results between their study and ours.

Additionally, D. Lomas et al. (2013) reported in their controlled experiment involving a basic educational game that simpler problems increased engagement but resulted in a slower rate of learning. They had initially hypothesized that moderate levels of challenge would maximize engagement, following the inverted-U hypothesis. They attributed these unexpected findings to their game not aligning with the typical close game hypothesis proposed by Abuhamdeh and Csikszentmihalyi (2012), where the challenge is motivating when there is high uncertainty about winning or losing. Importantly, their study differed from ours as it involved a different learning task and lacked explicit feedback after each question, thus missing the integration of uncertainty about winning or losing. Thus, there may be no such thing as a universal optimal difficulty level or consistent applicability of the inverted-U shape hypothesis across all learning scenarios. Rather, optimal training conditions may depend on the learning domain, on the type of training, as well as on students' ability and characteristics.

Indeed, we found evidence for such variability in our own results, with optimal difficulty levels ranging from -1.55 (about 83% success rate) in hematology to -0.53 (about 64%) in physical

medicine and rehabilitation, with a mean of -0.952 (SD = 0.378) (see Table3.8). This may reflect the state of the current BNE question bank, with questions in different specialties being designed by different professors, leading to specialties with questions of different difficulty levels on average.

Furthermore, even within a given specialty, optimal difficulty levels depended on the final student's ability. Higher-ability students tended to have more pronounced negative curvature, thus showing greater sensitivity to the mean relative difficulty, while lower-ability students tended to have flatter response curves. The interpretation is not straightforward, but this may in part be due to some lower-ability students responding at random regardless of question difficulty, thus being less sensitive to this factor.

Furthermore, in all but 4 specialties (out of 13), the optimal difficulty level for the best students was lower than that for the lower-ability ones. This may seem paradoxical, but one should recall that this is *relative* question difficulty (absolute question difficulty - online student ability). Thus, to take digestive surgery in Figure 3.2 as an example, an optimal difficulty level of -1 for students with an ability of 2 still represents a much greater absolute question difficulty (1) than an optimal difficulty level of 0 for students with an ability of -2 (-2). So it is the case that better students benefit from more difficult questions (in an absolute sense), but the optimal difference between question difficulty and student ability seems to decrease with ability. Why the opposite trend is observed in 3 specialties (hematology, hepatogastroenterology, and psychiatry) is not clear, and would first warrant replication before any attempt at an interpretation.

In terms of implications, our findings emphasize the importance of customizing practice exercises to individual needs. While many systems achieve this adaptation by dynamically adjusting difficulty levels based on past performance or by externally determining an optimal level, they often lack a firm empirical grounding. In the BNE system, it should now be possible to enhance learning outcomes by selecting questions at precisely the optimal level given a student's current ability in a given medical specialty.

Our study differs from existing research in several notable respects. Firstly, unlike prior studies that often focus solely on simple declarative knowledge, our platform encompasses a comprehensive array of medical knowledge spanning a six-year medical university curriculum. This complexity is reflected in our final exam, which assesses learning outcomes with an entirely new set of questions designed to evaluate students' ability to apply medical knowledge to unfamiliar scenarios. This differs from many other studies that rely on repeated questions within a much narrower knowledge domain.

Secondly, while some studies utilize educational games, our platform is specifically designed for training purposes, with the primary goal being compulsory learning rather than entertainment. This distinction may be important, as factors influencing intrinsically motivated activities may differ from those affecting non-intrinsically motivated activities (Elliot et al., 1994; Qi Zhang et al., 2022). Additionally, our analysis involves an extensive training period lasting four months, a duration notably longer than many other studies in the field.

Finally, unlike similar studies in the field that rely on learning curves established by reference questions (Papoušek, Stanislav, et al., 2016b), we used learning outcomes that were measured

through the scores of students in a final exam. This offered an ecologically valid external validation of learning.

Overall, all these factors contribute to the robustness and uniqueness of the present study.

4.1 Limitations

There are several notable limitations in our current study. Firstly, it is important to acknowledge that our findings are specifically tied to medical learning within the context of the chosen educational platform. The observed inverted U-shaped relationship for optimal difficulty, as demonstrated in our study, adds to the existing literature documenting this phenomenon across various disciplines, including digital gaming (Abuhamdeh and Csikszentmihalyi, 2012), online chess (Abuhamdeh, Csikszentmihalyi, and Jalal, 2015), and literacy training (Ronimus et al., 2014). Furthermore, our findings contribute to the well-established evidence supporting the positive impact of personalizing the difficulty level of learning content based on students' behaviors and performance (Major et al., 2021), reinforcing its relevance in adaptive learning systems. This cross-disciplinary consistency suggests that our findings may have broader applicability. Of course, none of the specific results and predictions reported here can be straightforwardly translated to other learning contexts, since they are specific to medical training in this cohort, indeed they even vary by specialty, and they also depend on the specific set of questions available on the platform. Rather, our results suggest that similar modeling or experimental manipulation of relative question difficulty in other learning contexts may produce similar effects, and thus it may be possible to determine an optimal difficulty level for each student in each learning context. Still, further research is required to explore the generalizability of these results across diverse educational contexts and subject areas.

Secondly, while previous studies often employ controlled conditions (Papoušek, Stanislav, et al., 2016b) or allow users to freely select the difficulty levels (D. Lomas et al., 2013), this study could not experimentally manipulate relative question difficulty in a randomized controlled design, which would have been the most powerful approach to address our research questions. However, we were able to exploit the random drawing of questions to adopt a quasi-experimental design in observational data. This is an equally valid way to estimate causal effects when direct intervention is not possible, which is frequently used in epidemiology and in economics.

More generally, contrary to experimental settings, we had no control whatsoever over the students' learning experience. Students were entirely free to use this platform for training or not: some did so very intensively, others very little. They were also free to choose specialties and types of questions (isolated questions, progressive files) to train on. Most students likely had other learning opportunities outside the platform, and so to different degrees. We had no information whatsoever on training performed outside the platform, we only had access to a subset of their overall training experience. The one parameter that was not controlled by students was question difficulty, whose effect we were, therefore, able to estimate. However, our estimation of relative question difficulty only concerns the part of their training that students did on the platform and ignores the difficulty of training done outside the platform. Thus, our evaluation of the effect of relative question difficulty must be underestimated.

The last limitation is the lack of engagement analysis in our study, due to constraints within our dataset. Existing literature suggests that difficulty levels may impact learning and engagement differently (Papoušek, Stanislav, et al., 2016b). Thus, incorporating engagement analysis could provide valuable insights into the relationship between difficulty, learning outcome, and student engagement.

4.2 Perspectives

The current study and its identified limitations prompt new research questions. Firstly, it would be desirable to re-assess our research questions using much more controlled settings, with learners' training experience entirely documented, and ideally using a randomized controlled design to experimentally manipulate relative question difficulty.

Building upon the identification of optimal difficulty levels for various medical specialties and student abilities, it would seem logical to implement an adaptive learning algorithm within the educational platform, which would dynamically present questions at the appropriate difficulty level for each learner in each specialty. It would then be important to assess whether overall learning outcomes improve compared to the non-adaptive version of the platform.

While our study offers insights into the relationship between difficulty and learning outcomes, it leaves unanswered questions regarding student motivation and engagement. Recent research has yielded mixed findings in this area: while one study found no clear effect of difficulty-skill balance on engagement (Cutting et al., 2023), another (Israel-Fishelson et al., 2021) reported a positive association between task difficulty and micro-persistence—defined as the tendency to successfully complete individual tasks, a nuanced measure of engagement. These mixed results underscore the need for further investigation into how difficulty influences engagement, aiming to achieve a balanced approach that offers exercises not too hard, and not too boring.

Furthermore, a recent study (J. D. Lomas et al., 2017) discovered that when users were allowed to choose their own difficulty levels, moderately challenging tasks were the most motivating. Conversely, when difficulty was predetermined, the easiest tasks were found the most motivating. This suggests exploring the possibility of leaving learners the choice of training difficulty level. It would then be interesting to evaluate whether this has the intended effect on motivation, and furthermore whether this also has a positive effect on learning.

5 Conclusion

This study provides significant insights into the effects of training difficulty on learning outcomes, supporting the Inverted U-shape Hypothesis, which posits that there is an optimal level of difficulty that maximizes learning. Furthermore, it demonstrates that this optimal level of difficulty varies according to the subject matter and the learner's proficiency in that subject. These insights have important implications for the education and learning analytics communities, advocating for the adjustment of training difficulty in an adaptive and personalized way to ensure that all learners encounter optimally challenging content, tailored to their individual learning curves and the specific demands of the subjects they are studying.

Chapter 4

A Meta-analysis of the Impact of Feedback Timing on Learning Outcomes

This section constitutes the following manuscript:

Kandemir, E. N, Esposito, E., Gurgand, L., & Ramus, F.(2025).
A Meta-analysis of the Impact of Feedback Timing on Learning Outcomes. (under review)

Both the preregistration for this meta-analysis and the data and analysis scripts are available on the Open Science Framework (OSF):

- Preregistration
- Data and analysis scripts

Contents

ABSTRACT This study investigates the effect of feedback timing on learning outcomes in computer-assisted learning environments. Three key aspects were examined: (a) the overall effectiveness of immediate versus delayed feedback, (b) the influence of varying definitions of feedback delay on learning outcomes, and (c) study characteristics and feedback features that may account for inconsistencies in prior findings. A systematic and quantitative meta-analysis was conducted on 51 studies published from 1988 to 2024, analyzing 160 effect sizes through meta-regression with robust variance estimation. Results indicate that feedback timing does not significantly influence learning outcomes on average ($g = 0.03$, 95% CI $[-0.07, 0.14]$, $p = 0.518$). However, substantial heterogeneity between studies suggests that the effect is moderated by key factors. Moderator analyses reveal that educational level, learning domain, post-test task type, and response time constraints significantly influence the effectiveness of feedback timing, providing a partial explanation for inconsistencies across studies. However, the partial confounding of moderators across studies makes it impossible to fully identify their respective influence on feedback timing effects. These findings suggest that while feedback timing alone is not a decisive factor, its impact is context-dependent. Potential explanations, practical implications, and limitations of the findings are discussed, along with directions for future research.

1 Introduction

Feedback refers to information provided regarding learners' performance, aimed at assisting them in confirming, expanding, or modifying their stored knowledge (Sadler, 1989; D. L. Butler et al., 1995; Hattie and Timperley, 2007). The critical impact of feedback on education is well-documented through extensive meta-analyses (Wisniewski et al., 2020; Hattie and Clarke, 2018; Hattie and Timperley, 2007).

With the rapid integration of technology into education, digital learning environments are now offering advanced feedback mechanisms that can enhance learning outcomes more effectively than traditional methods, although the degree of this impact varies based on the specific characteristics of the feedback (Cai et al., 2023; Mohamed et al., 2020; Van der Kleij, Feskens, et al., 2015; Mohsen, 2022; Azevedo et al., 1995). One notable feature of digital learning environments is the systematicity and precision with which they can deliver feedback. Unlike traditional educational methods that relied mostly on teacher correction, digital learning environments can provide feedback almost immediately after a response is recorded. This capability has significantly advanced the scope of research into feedback timing.

Research on the impact of feedback timing dates back to the 1980s, with the seminal meta-analysis by Kulik et al. (1988), which investigated whether receiving feedback immediately or after a delay is more effective for learning. Although few studies at the time used digital tools, the findings indicated that immediate feedback generally had small to moderate positive effects, compared to delayed feedback. However, subsequent research has provided mixed results, leaving the role of feedback timing in digital learning environments unclear. To the best of our knowledge, no meta-analysis has specifically focused on studies that directly compared immediate and delayed feedback since Kulik et al. (1988). This study seeks to address this gap through a comprehensive meta-analytic review of empirical research.

1.1 Theoretical Perspectives on Feedback Timing

In their review, Shute (2008) classified feedback timing in educational settings into two categories: immediate and delayed. Immediate feedback is defined as feedback given right after a learner completes a task or a problem-solving step. Delayed feedback, on the other hand, is provided after an extended period, ranging from several hours to several days, or even up to a week following task completion. This categorization naturally leads to a debate over whether it is more effective to provide feedback immediately or after a delay.

The primary justification for favoring immediate feedback over delayed feedback is rooted in behaviorist psychology, which suggests that feedback (or reinforcement) must be immediate to effectively influence learning (or conditioning) (Hull, 1952; Saltzman, 1951; Skinner, 1965). Additionally, cognitive load theory (Sweller, 2011) also supports immediate feedback by indicating that working memory's limited capacity is best managed when learners receive detailed feedback in digestible amounts, thus avoiding cognitive overload.

On the other hand, three theoretical frameworks commonly advocate for the advantages of delayed feedback: the interference-perseveration hypothesis, the dual-trace hypothesis, and

the attention-based account. The interference-perseveration hypothesis (Kulhavy and R. C. Anderson, 1972) argues that delayed feedback is more effective for correcting errors because the delay allows initially incorrect responses to fade from working memory, therefore minimizing confusion and enhancing the acquisition of the correct response. This theory assumes that errors hinder learning, although subsequent research indicates that generating errors can actually facilitate learning (A. C. Butler, Fazio, et al., 2011; Metcalfe and J. Xu, 2018).

The dual-trace hypothesis (Kulik et al., 1988; Clariana et al., 2000) proposes that delayed feedback fosters learning by offering two distinct encoding opportunities: one at the time of the initial response and another during feedback, effectively reinforcing correct responses and correcting errors.

The attention-based account (Phye et al., 1989; Kulhavy and R. C. Anderson, 1972) suggests that delayed feedback is more effective because learners are more likely to pay attention to and engage with feedback that is separated from the learning event, giving them more time to process and understand the feedback.

Finally, the distributed practice effect offers an alternative, memory-based explanation for benefits observed with delayed feedback. According to this perspective, delayed feedback does not enhance learning because of the delay itself, but because it creates an additional spaced-retrieval opportunity. Research on the spacing effect shows that distributing retrieval attempts over time strengthens memory retention more effectively than massed practice (Nicholas J Cepeda et al., 2009; Nicholas J. Cepeda et al., 2008). Thus, improvements following delayed feedback may not reflect a true effect of feedback timing per se, but rather the benefits of spaced retrieval and memory consolidation, particularly for initially correct responses (A. C. Butler, Karpicke, et al., 2007; Metcalfe, Kornell, et al., 2009).

1.2 Previous Meta-analyses and Reviews

Immediate feedback is often preferred over delayed feedback by educators and particularly by students (Lefevre et al., 2017; Van der Kleij, Eggen, et al., 2012; Mullet et al., 2014), under the assumption that it facilitates better learning outcomes. This preference has influenced the design of many online courses, learning platforms, and applications, which are structured to provide immediate feedback (e.g., Kahoot!, Duolingo, Quizlet, ALEKS). However, a detailed examination of the literature reveals mixed results.

The meta-analysis by Kulik et al. (1988), included 53 studies from 1966 to 1988 specifically comparing these two feedback timings. They found that immediate feedback generally outperformed delayed feedback in classroom studies, while in laboratory settings, delayed feedback is more effective. The meta-analysis also explored how the delay is defined —either after each item or after the entire test—affected outcomes, and revealed nuanced differences in the effectiveness of these two different definitions.

Other meta-analyses and reviews have more generally investigated the impact of feedback on learning outcomes, using feedback timing as a moderator. For instance, Azevedo et al. (1995) assessed the influence of computerized feedback on learning by synthesizing 22 studies published until 1992. They found that immediate feedback significantly enhanced students' aca-

demic achievement compared to delayed feedback. A more recent analysis by Van der Kleij, Feskens, et al. (2015) examined feedback effects in computerized settings across 40 studies from 1968 to 2012. They categorized feedback timing into immediate and delayed, defining delayed feedback as any feedback "not delivered immediately after completing each item". Their findings indicated that immediate feedback was particularly effective for lower-order learning tasks, whereas delayed feedback seemed more suitable for higher-order tasks, though they reported no significant interaction. Swart et al. (2019) conducted a meta-analysis of 60 studies from 1962 to 2016, focusing on the feedback effect on reading performance. Their moderator analysis revealed that feedback was most effective when provided immediately after a learning session rather than during the session itself. Furthermore, in their review, Jaehnig et al. (2007) examined 33 studies from 1964 to 2004, with various feedback types and timings. They found no significant differences in learning outcomes between immediate and delayed feedback.

Finally, a very recent review M. Xu et al. (2023) investigated 20 empirical studies on second language learning from 2006 to 2021, which directly compared immediate and delayed feedback. The results suggested that immediate feedback was often more effective or equally effective compared to delayed feedback, with variations likely due to differences in communication modalities, feedback explicitness, and the specific timing of delays. Thus, the role of feedback timing may not be constant across studies but may well depend on specific learning contexts and study characteristics.

1.3 Possible Moderators of feedback timing effects

Previous review studies on feedback effectiveness have primarily focused on core characteristics such as timing and type (Van der Kleij, Feskens, et al., 2015; Van der Kleij, Eggen, et al., 2012). More recently, attention has expanded to include additional dimensions like presentation format (Donkin et al., 2019) and adaptability to learner needs (Janelli et al., 2021; Lim et al., 2021). Research also suggests that factors such as learners' prior knowledge, learning environment, and outcome type can influence feedback effectiveness in digital learning contexts (Swart et al., 2019; Van der Kleij, Feskens, et al., 2015).

Drawing from this literature, we identified key moderators that may explain inconsistent findings regarding feedback timing. These fall into two categories: (1) experimental design features—such as feedback type, how delay is defined, the actual time gap between feedback conditions, and task context; and (2) learner-related factors—including educational level, subject domain, and task complexity.

Feedback Type Feedback type is a key characteristic of educational feedback and has been widely examined in meta-analyses and reviews. Shute (2008) categorized feedback into two types: simple and elaborated feedback. Simple feedback, comprising knowledge of results (KR) and knowledge of correct response (KCR), provides basic right-wrong information and, in the case of KCR, the correct answer. Elaborated feedback (EF), however, not only indicates correctness but also provides detailed explanations and guidance for improvement.

While Brummer et al. (2024) reported no significant effects of feedback type on learning, Cai

et al. (2023) found that certain types of EF are significantly more effective than simpler forms. These findings align with Van der Kleij, Feskens, et al. (2015) and Mertens et al. (2022) who noted a consistent trend across studies where richer feedback content more effectively enhanced learning outcomes. Moreover, Swart et al. (2019) explored the interaction between feedback type and timing and revealed that EF and KCR are more effective when provided after, rather than during reading.

Individual studies examining the interaction between feedback type and timing also offer diverse insights. For example, Smits et al. (2008) and Corral et al. (2021) found no significant interaction when comparing EF with KCR. However, Taxipulati et al. (2021) identified a significant interaction, highlighting that learners responded more attentively to adaptive feedback when it was provided immediately rather than delayed. Furthermore, Roper (1977) suggested combining immediate KR with delayed EF, suggesting that quick correctness verification followed by a period for reflection best enhances learning.

This variability underscores the necessity of considering both feedback type and timing in educational research to better understand the inconsistent findings in prior feedback timing research.

Definition of Feedback Delay and Delay Difference Between Immediate and Delayed Feedback Immediate feedback typically occurs right after a student responds to a question. The definition of delayed feedback, however, can span from a few hours to a week following task completion (Shute, 2008; Dempsey et al., 1988). With advancements in technology, particularly in computer-based environments, Van der Kleij, Feskens, et al. (2015) recommends considering any feedback that is not immediate in such contexts as delayed. However, as delivery methods diversified, two distinct ways to define feedback delay have emerged. Some studies in computer-based environments define delay by the number of intervening items before feedback is given — ranging from a few items to the end of an entire task — while others define it by time, providing feedback after a single item or after completing the entire task, but with delays spanning several seconds, hours, or even a week.

This variability clearly shows that the definition of feedback delay may be an important contributor to inconsistent findings in research on the impacts of immediate versus delayed feedback (Kulik et al., 1988; Mory, 2013; M. Xu et al., 2023). This suggests that merely categorizing feedback as immediate or delayed, as seen in previous meta-analyses, may not sufficiently capture the nuances affecting learning outcomes. Instead, ensuring that studies are comparable requires aligning them on the same delay scale—whether in terms of items or time. Then, the precise delay difference between feedback defined as immediate and as delayed in each study can be used as moderator, allowing for a consistent and reliable comparison across different studies.

Learning Task Context The learning process can differ significantly depending on whether the task is designed purely for research purposes, or if it is integrated into a real curriculum, directly benefiting learners' educational or professional goals. Although not explicitly catego-

rizing these two contexts, Kulik et al. (1988) explored the effects of feedback timing in both classroom and laboratory settings. They concluded that in classroom studies, immediate feedback tends to enhance learning more effectively, whereas in laboratory studies, delayed feedback often proves superior. This disparity may be attributed to differences in motivation, cognitive load, and relevance to real-world applications, which are also relevant to the distinction between experimental and curriculum-based learning tasks. Therefore, the context of the learning task should be thoroughly investigated as a potential moderator of the feedback timing effect.

Education Level The impact of immediate versus delayed feedback on learning may differ by educational level, as there are cognitive and engagement differences between students at different levels (Swart et al., 2019). While a meta-analysis by Van der Kleij, Feskens, et al. (2015) found that computerized feedback was more effective for college students than for those in primary and middle school, the role of educational level as a moderator of feedback timing has not been extensively tested since Kulik et al. (1988)'s findings, which found no significant effect. Considering the advancements in computerized education and additional research conducted since 1988, education level may be a critical moderator in the effectiveness of feedback timing across studies.

Learning Domain The optimal timing for correcting erroneous knowledge may vary across different learning domains due to their unique characteristics. Research by Van der Kleij, Feskens, et al. (2015) suggested that computer-based feedback significantly enhanced math performance but had minimal impact on language learning. However, this finding was influenced by the predominance of elaborated feedback in math in the studies reviewed and requires cautious interpretation. Furthermore, the results of Mertens et al. (2022) did not show larger effect sizes in mathematics compared to other domains but did reveal that the effectiveness of specific feedback types varied by learning domain. These observations highlight the need to examine the role of the learning domain as a moderator of feedback timing effects.

Training Task Complexity Training task complexity within a learning domain may also influence the effectiveness of feedback timing. Lower-order tasks, involving knowledge, comprehension, recall, understanding, and reproduction, differ from higher-order tasks, which require application, analysis, synthesis, evaluation, and strategic thinking (Bloom et al., 1964). According to Mason et al. (2001), immediate feedback is more beneficial for lower-achieving students regardless of task level, whereas higher-achieving students benefit more from delayed feedback, particularly with higher-order tasks. Shute (2008) concurs, suggesting that immediate feedback is optimal for lower-order learning outcomes, while higher-order outcomes benefit from a delay. Similarly, Van der Kleij, Feskens, et al. (2015) observed that immediate feedback enhanced lower-order task performance, but found no significant interaction between learning outcome levels and feedback timing. Therefore, training task complexity should be considered a potential moderator of the effectiveness of immediate versus delayed feedback.

1.4 Limitations of Existing Meta-Analyses on Feedback Timing

Although previous meta-anlyses shed light on the effects of feedback timing on learning outcomes, they present limitations, especially in digital learning environments. Firstly, the variability in the definitions of "immediate" and "delayed" feedback across studies introduces inconsistencies and complicates the interpretation of meta-analytic results. Only Kulik et al. (1988), which included a small subset of studies using computer-based feedback, attempted a more nuanced differentiation between delays by categorizing them as either post-item or post-test. Their findings indicated that when delayed feedback was given only a few seconds before the next item, immediate feedback was significantly more effective. However, when feedback was postponed until after an entire test, this advantage disappeared, with delayed feedback even leading to slightly better performance. Despite this attempt at differentiation, further refinement is needed to capture the spectrum of feedback timings used in digital learning environments.

Secondly, there is a noticeable lack of recent meta-analyses that directly compare immediate and delayed feedback within the context of modern digital learning environments. Earlier reviews like those by Azevedo et al. (1995), Van der Kleij, Feskens, et al. (2015), and Swart et al. (2019) primarily used feedback timing as a moderating factor in meta-analyses of feedback effects, which may obscure specific effects attributable to timing. The most recent direct comparison of immediate and delayed feedback goes back to Kulik et al. (1988) and therefore misses nearly four decades of studies on feedback timing.

1.5 The present Study

Given the inconsistent findings regarding the optimal timing of feedback, limitations in previous meta-analyses, rapid technological advancements in education, and the significant expansion of empirical research on feedback timing, there is a pressing need to re-examine and synthesize the literature on feedback timing within digital learning settings. This meta-analysis aims to clarify the effectiveness of different feedback timings by focusing specifically on studies that directly compare immediate feedback with delayed feedback. The key research questions are:

- **RQ1:** What is the difference between the effects of immediate and delayed feedback on learning?
- **RQ2:** How do varying definitions of "immediate" and "delayed" feedback influence the reported effectiveness of these feedback timing effects on learning outcomes?
- **RQ3:** What other factors moderate the effects of immediate versus delayed feedback on learning outcomes?

2 Methods

Our meta-analysis was conducted in accordance with the PRISMA guidelines (Page et al., 2021) to address the defined research questions. The methodology included pre-registered[1]

[1]Pre-registration available at OSF repository

inclusion and exclusion criteria, a structured search protocol, a systematic screening protocol, data extraction protocols, and detailed statistical methods.

2.1 Inclusion and exclusion criteria

In line with previous review studies on the effects of feedback timing on learning, and with our research objective to calculate the aggregate effect of immediate versus delayed feedback on learning, we defined the following inclusion criteria:

(a) Publication Type: Studies must be published in peer-reviewed scientific journals, unpublished doctoral dissertations, or in conference proceedings.

(b) Publication Year: Studies must be published after 1988. This cutoff was set to include the corpus of studies not reviewed in the last major meta-analysis (Kulik et al. (1988)).

(c) Publication Language: Publications must be in English, Turkish, or French, covering a broad spectrum of educational research within our linguistic capabilities.

(d) Educational Context: Studies must involve meaningful learning processes, such as acquiring new knowledge, skills, or concepts, regardless of whether they occur in formal education settings or controlled experimental environments designed to promote learning.

(e) Learning Environment: Studies should be conducted in computer-assisted settings to standardize feedback delivery and constrain the delay definition. This includes:

- Studies explicitly conducted within a computer-assisted environment.
- Studies utilizing digital communication platforms (e.g., Skype, Zoom) for experimental procedures.
- Experiments conducted in labs using digital tools or software.
- Any setting where the training and feedback components are delivered through computer systems, irrespective of the format of assessments (digital or directly by an instructor).

(f) Participant Type: Participants must be typical learners without special educational needs to ensure consistency in learning capabilities across studies. Studies are assumed to qualify whenever no specific disorders or disabilities are mentioned.

(g) Feedback Definition: The feedback operationalized in the study should comply with the following criteria:

- Feedback must be provided on an individual basis, directly related to participants' responses or performances.
- Feedback timing should be clearly defined as either immediate or delayed. Studies using alternative terminology (e.g., synchronous/asynchronous) that logically imply a comparison between immediate and delayed feedback also meet this criterion.

- Although exact times or item counts are not mandatory, the description should categorize the feedback as item-based or time-based, and providing an indication of the implemented delays.

(h) Independent Variable: Studies must directly compare the effects of immediate versus delayed feedback, addressing our central research question.

(i) Dependent Variable: The study should focus on learning outcomes as a primary measure of success. Learning outcomes must be measurable, such as post-test scores or skill acquisition levels, to ensure an objective assessment of feedback effectiveness.

(j) Study Design: Studies must employ a randomized controlled experimental design to minimize potential biases and eliminate confounding variables. This includes studies where participants are randomly or pseudorandomly assigned to conditions that are identical except for the timing of the feedback, with no obvious confounding factors.

(k) Statistics: Studies must provide sufficient statistical data (e.g., means, standard deviations, t-values, F-values) to allow for the extraction or computation of effect sizes.

Deviations from pre-registered inclusion criteria: The above inclusion and exclusion criteria are those that were pre-registered, with four modifications: (1) We added a criterion related to the educational context to exclude studies focused on non-semantic learning —tasks without meaningful content or educational value, such as perceptual discrimination—, as these were less relevant to our main purpose. (2) We refined the feedback definition criterion after encountering studies where feedback was based on rankings or behaviors not directly linked to task responses, or where feedback was summarized (e.g., number of correct responses) rather than tied to specific answers. (3) We refined the definition of a computerized environment to include specifically studies where training and feedback are delivered via computer systems, regardless of whether pretests and posttests were computerized, thus focusing on the importance of the training phase. (4) We initially required detailed reporting of delayed feedback timing for the purpose of moderator analyses, but we relaxed it as it was not critical for calculating the general pooled effect size. This allowed the inclusion of two additional studies (E. Lavolette et al., 2015; Dunbar et al., 2017) which did not provide specific delay details but met all other inclusion criteria.

2.2 Search Protocol

To identify studies for inclusion in our meta-analysis, we conducted systematic searches across three major databases: Web of Science, Scopus, and ERIC. The comprehensive search terms are detailed in the Appendix (see Appendix 0.1), and the dataset of retrieved papers is available on the OSF project page.

To ensure the efficacy of our search strategy, we tested it against a core set of seminal papers (Attali and Kleij, 2017; Van der Kleij, Eggen, et al., 2012; Metcalfe, Kornell, et al., 2009; Nakata, 2015), which are pivotal in this research area. This validation step helped to confirm the

inclusivity and accuracy of our search terms, allowing for necessary adjustments. Our database search yielded 6,422 papers. After removing 1,322 duplicates, we were left with 5,100 papers.

Additionally, we set up alerts in both Web of Science and Scopus to notify us of new publications relevant to our search terms, until December 6, 2024, when we froze the database. This feature ensured that we were informed weekly of the latest studies, allowing for timely inclusion in our analysis. We have also examined the reference lists of eligible studies and relevant review papers identified during the initial search. This reverse search process and the inspection of the "relevant studies" section in Web of Science led to the inclusion of 8 additional records.

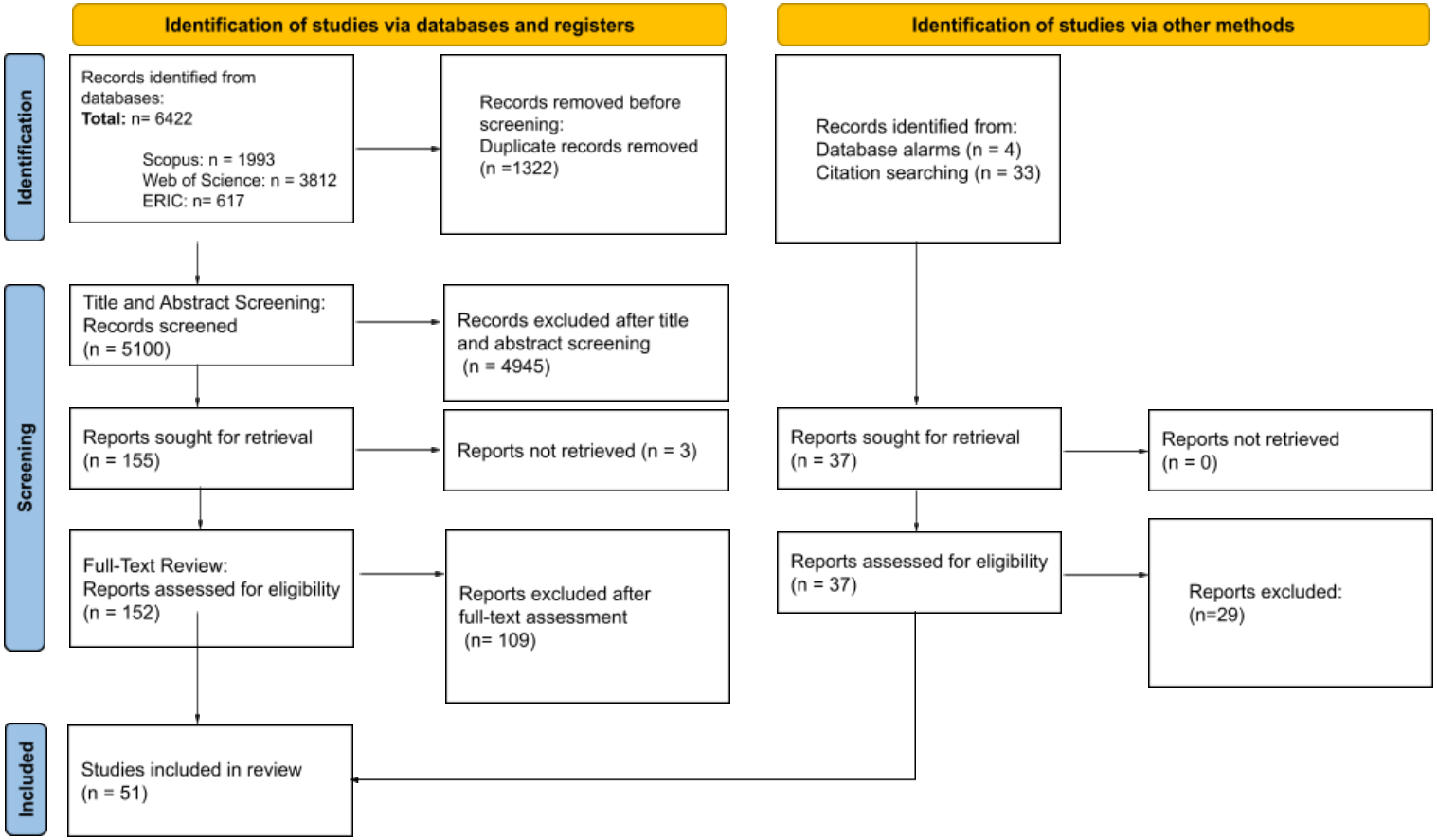


Figure 4.1: PRISMA flow diagram based on the PRISMA 2020 model.

For an overview of the search steps and the corresponding numbers of articles at each stage, refer to the PRISMA diagram in Figure 4.1.

2.3 Screening Protocol

The selection of included studies involved a methodical two-stage screening process, following initial retrieval from our database searches. These stages are summarized in Figure 4.1.

The initial screening was based solely on information available in the title and abstract, focusing on criteria such as study focus and methodology. The first author conducted this screening, which led to the exclusion of 4,945 papers. To verify the accuracy and consistency of this screening, approximately 10% of these records (n = 500) were independently reviewed by a second screener. Inter-rater reliability was substantial, with a Kappa coefficient of 0.742 (95% CI [0.725, 0.759]), indicating a high level of agreement (Landis et al., 1977). Cases of disagreement (N = 8) were discussed and resolved jointly. In 7 out of these 8 cases, the final decision matched the initial judgment of the first screener.

The remaining 155 papers underwent a full-text review against all inclusion and exclusion criteria, along with a quality assessment conducted using the tool proposed by Kmet (2004). This tool, which evaluates 14 different criteria, is designed to assess the rigor of quantitative research through the adequacy of research questions, study design appropriateness, and thoroughness in result reporting. Three studies (Schroth, 1995; Schroth, 1992; Schroth and Lund, 1993), contributing a total of 21 effect sizes, were excluded during this quality assessment stage due to critical methodological confounds. Although they met the core inclusion criteria, their training procedures required participants to reach a performance threshold (e.g., 9 out of 10 correct) rather than undergoing a fixed number of trials. As reported by the authors, this resulted in substantially more training for participants in the delayed feedback condition. Thus, while they concluded that delayed feedback enhanced transfer, this advantage is confounded by unequal training exposure, making it unclear whether the effect is attributable to feedback timing or additional practice.

After excluding a total of 109 papers in this full-text review stage and an additional 3 that could not be retrieved, 43 papers were found eligible for data extraction. Approximately 13% of the 155 papers (n = 20) were independently reviewed to assess reliability, achieving a Kappa value of 0.70 (95% CI [0.510, 0.890]), indicating a high degree of consistency (Landis et al., 1977). A detailed record of reasons for exclusion at each screening stage is available on the OSF project page.

With the addition of 8 papers found through other methods, we ended up with 51 studies from which the effect size was extracted and moderator variables were coded. A full list of the 51 included studies, along with coded preregistered moderator variables, is presented in Table 6 (see Appendix 0.2).

2.4 Data Extraction

2.4.1 Extraction of statistics for calculating effect sizes

The procedures for calculating effect sizes is detailed in the Statistical Methods section. To ensure accurate computation, we followed several guidelines during the initial coding of relevant statistics:

(a) To ensure consistency, the standardized mean differences were always coded with delayed feedback as the reference group. Thus, a positive mean difference indicates that immediate feedback is more effective than delayed feedback, and a negative effect size suggests the opposite.

(b) Following the recommendations by Lipsey (2001), priority was given to directly available statistics such as means and standard deviations. When these were not available, we used other statistics (e.g., t-tests, regression coefficients, ANOVA results) and transformed them according to established formulas.

(c) When studies reported statistics for multiple conditions, we generally coded them separately unless the conditions were not relevant to our research questions or moderators.

(d) If studies reported both raw and adjusted learning outcomes (adjusted for prior knowledge, usually using pre-test results), the adjusted outcomes were coded. Otherwise, statistics were based on raw data. This was also recorded as a moderator in order to test whether this made a difference.

The initial coding of these statistics was conducted by the first author. To validate the reliability of this coding, two additional coders independently coded 10% of the included studies each, totaling 20% of the studies being double-coded. These coders were also tasked with calculating effect sizes based on their extracted statistics. Discrepancies between primary and secondary coders were resolved through discussion, with no significant issues affecting the final dataset. The correlation of effect sizes between the double-coded studies was $r = 0.96$ for 32 effect sizes.

2.4.2 Coding of Study Characteristics

In order to address inconsistencies in previous findings and conduct moderator analyses, we coded additional characteristics for each included study. The pre-registered moderators were the following:

Delay Unit: Feedback delay unit was categorized into two types: item-based and time-based. In an item-based delay, the interval between an answer and its feedback is determined by the number of intervening items (e.g., 3 items). In a time-based delay, the interval is determined by the amount of time that elapses between an answer and the corresponding feedback (e.g., 10 seconds).

Delay Difference: The delay difference between immediate and delayed feedback, in seconds or in the number of items, depending on the delay unit.

Feedback Type: Drawing on the categorization proposed by Shute (2008), feedback was classified into four primary types: Verification, Correct Response, Try Again, and Elaborated. For simplicity, Verification and Correct Response were combined into a single category termed "Simple Feedback Type."

Education Level: Categorized into Primary, Secondary, Tertiary, and Adult Education levels.

Learning Domain: Studies were classified into specific learning domains:

- Text Memorization: Tasks involving memorization of text for later recall.
- Language Learning: Tasks related to language skills, including vocabulary acquisition and memorization.
- Reading Comprehension: Focused on understanding and analyzing text.
- STEM: Focused on science, technology, engineering, and mathematics.
- Social Sciences: Includes studies on social behaviors, cognition, and societal issues.
- General Cognitive Skills: Encompasses cognitive tasks outside of language or STEM categories, such as probabilistic learning and general memory tasks.
- Mixed: Studies spanning multiple domains or interdisciplinary approaches.

Training Task Complexity: Using Bloom et al. (1964) taxonomy, tasks were classified as Lower-Order (involving knowledge, comprehension, or recall) and Higher-Order (requiring application, analysis, synthesis, evaluation, or extended thinking). Studies encompassing both were categorized as Higher-Order.

Prior Knowledge Adjustment: Indicates whether the effect sizes were adjusted for participants' prior knowledge and skills, as assessed through pretests administered before training.

Learning Task Context: Experimental Tasks: Designed specifically for research, often detached from practical educational applications. Curriculum-Based Learning Tasks: Integrated within an actual curriculum, enhancing learners' educational or professional objectives.

In addition to the pre-registered moderators, several other study characteristics were coded. The impact of these additional characteristics was also explored in our moderator analysis. Below is a list of these characteristics along with their respective coding schemas:

- Participant Design: within-subject / between-subjects.

- Learning Task Type: classified as arbitrary / non-arbitrary, based on the meaningfulness of the learned associations. Arbitrary tasks involve learning associations that do not have inherent semantic connections (e.g., pairing unrelated words or symbols), while non-arbitrary tasks involve meaningful or conceptually related content (e.g., reading comprehension, solving math problems).
- Post-test Item Similarity: identical to / different from training items.
- Feedback Dependency on Errors: feedback provided only for incorrect responses / systematically.
- Feedback Answer Reminder: Indicates whether the feedback included a reminder of the initial answer or not.
- Number of Training Items: The total count of items used during training sessions.
- Post-test Task: memory retrieval/ knowledge application
- Time Limitation for Responses: time-limited / no time limit for answering during tasks.
- Retention Interval Difference (between Last Feedback and Post-test): The time difference, in seconds, between the last feedback and the post-test between immediate and delayed feedback conditions. Calculated as the retention interval of delayed feedback subtracted from that of immediate feedback.
- Publication Year: The year the study was published.

Deviations from Preregistered Coding Procedure: Our analysis deviated from the preregistered coding procedure by combining the 'Verification' and 'Correct Response' feedback types into a new category named 'Simple Feedback Type.' This adjustment, not specified in the preregistration, was made to simplify the analysis and clarify impact distinctions between simpler and more complex feedback types. Additionally, this grouping was implemented to increase the number of effect sizes per group, enhancing the statistical robustness of the moderator analysis.

2.5 Statistical Methods

2.5.1 Calculation of effect sizes

In this meta-analysis, each effect size indicates the standardized difference in learning outcomes between groups under different feedback timing conditions, typically assessed using post-test scores.

To calculate the effect size we used the the data coded in the data extraction section and followed these formulas:

Cohen's d was calculated by:

$$d = \frac{M1 - M2}{S}$$

where $M1$ and $M2$ are the means of the immediate feedback and delayed feedback groups, respectively, and S denotes the pooled standard deviation. The pooled standard deviation was determined differently based on the study design:

For within-subject designs:

$$S = \sqrt{\frac{s1^2 + s2^2}{2}}$$

where $s1$ and $s2$ are the standard deviations for each of the two feedback timing conditions.

For between-subject designs:

$$S = \sqrt{\frac{(n1-1)s1^2 + (n2-1)s2^2}{n1+n2-2}}$$

where $s1$ and $s2$ are the standard deviations for each of the two feedback timing groups and $n1$ and $n2$ are the numbers of participants in each group.

The standard error of the effect size ($d.se$) was again calculated based on the study design.

For within-subject designs:

$$d.se = \sqrt{\frac{2(1-r)}{n} + \frac{d^2}{2n}}$$

For between-subject designs:

$$d.se = \sqrt{\frac{n1+n2}{n1 \times n2} + \frac{d^2}{2(n1+n2)}}$$

Here, r denotes the within-subject correlation between the conditions and was assumed to be 0.5 if not reported. As is standard in meta-analyses, standard errors were based on d, which was then adjusted to Hedges' g to correct for small sample bias:

$$g = d \times \left(1 - \frac{3}{4N-9}\right)$$

where N is the total sample size encompassing all conditions. All reported effect sizes use Hedges' g.

In cases where the necessary statistics were not directly accessible, alternative statistical measures such as t values, univariate F values, or χ^2 statistics along with sample sizes were employed. These were converted into effect sizes using established conversion formulas from the meta-analysis methodology literature (see Lipsey (2001) for detailed formulas).

2.5.2 Outlier Detection and Publication Bias

We assessed potential outliers in the final effect sizes (Hedges' g) by identifying any deviations exceeding three standard deviations from the mean effect sizes of several groups. This method aims to pinpoint effect sizes that might disproportionately influence the summary statistics or introduce bias. This approach was adapted from Van der Kleij, Feskens, et al. (2015), who explored feedback effects in computer-based learning environments.

To evaluate publication bias in our meta-analysis, we employed visual inspection via a funnel

plot to identify differences between estimates from small and large studies. Additionally, we conducted Egger's regression test (Egger et al., 1997), which assesses data asymmetry and offers a robust indication of potential biases within the included studies.

2.5.3 Computation of Weighted Mean Effect Sizes

All analyses reported in this study used Hedges' g for effect size measurement. We used the Robust Variance Estimation (RVE) method to address the common issue of multiple and non-independent effect size estimates arising from the same study (Hedges et al., 2010). This approach is preferred over traditional random effects models, which are not suitable for meta-analyses involving dependent effect size estimates. The RVE method is particularly advantageous as it provides a more accurate estimation of standard errors, leading to narrower confidence intervals for the weighted mean effect sizes (Hedges et al., 2010).

We applied the RVE model using the R software package *robumeta* (Fisher et al., 2015). This package requires specifying the correlation (ρ) between within-study effects. We used the default setting of $\rho = 0.80$ for consistency. In order to evaluate the potential impact of ρ on both the mean effect size and the estimated between-study heterogeneity (T^2), we conducted sensitivity analyses. These analyses varied the correlation of within-study effect sizes from $\rho = 0.0$ to $\rho = 1.0$, allowing us to explore the robustness of our findings across different assumptions (Tanner-Smith et al., 2014).

2.5.4 Moderator Analysis

In accordance with our pre-registered analysis plan, we conducted separate meta-regressions for each moderator. These regressions were integrated into the multilevel meta-analytical models by including each moderator as a fixed effect.

2.5.5 Multiple Meta-Regression

To complement the individual analyses of moderators, we conducted a multiple meta-regression to assess the influence of multiple moderator variables while controlling for the effects of others (Rubin, 1992). We only included the moderators that demonstrated significant effects in the separate analyses. To enhance statistical power, some moderator categories were consolidated by grouping their categories.

Deviations from Preregistration The multiple meta-regression was not part of our pre-registration but was introduced because the significant moderators appeared to be confounded with each other, raising the question of which of those actually influenced the meta-analytic effects.

The complete set of analysis codes, along with the data utilized in this study, is accessible on the OSF project page.

3 Results

3.1 Descriptive Statistics

This meta-analysis includes 51 studies conducted from 1995 to 2024, collectively contributing a total of 160 effect sizes. On average, each paper reported approximately 3.14 effect sizes, with a maximum of 12 and a median of 2 effect sizes per study. Two papers, accounting for six effect sizes, originated from dissertations, while the remaining 49 papers, contributing 154 effect sizes, were sourced from peer-reviewed journal articles. The geographical distribution of studies and effect sizes is summarized in Table 4.1.

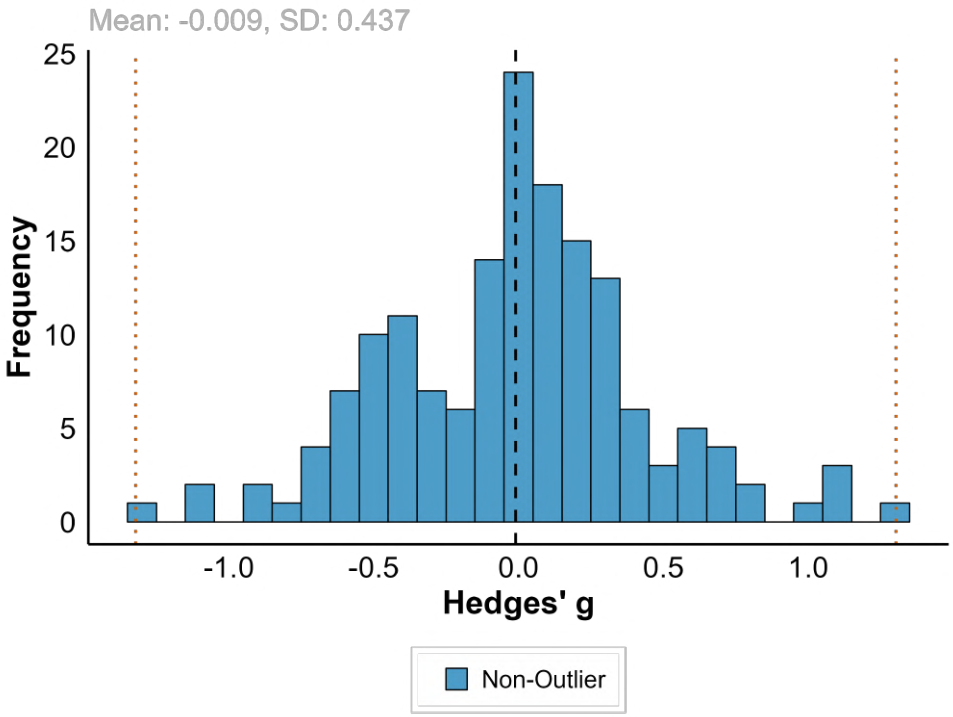
Table 4.1: Geographical distribution of studies and effect sizes.

Country	**Num. Study**	**Num. Effects**
United States	28	87
Japan	5	20
Spain	3	4
Netherlands	3	7
Germany	3	7
China	3	13
Australia	1	8
Taiwan	1	2
Saudi Arabia	1	3
Greece	1	2
Unspecified	2	4

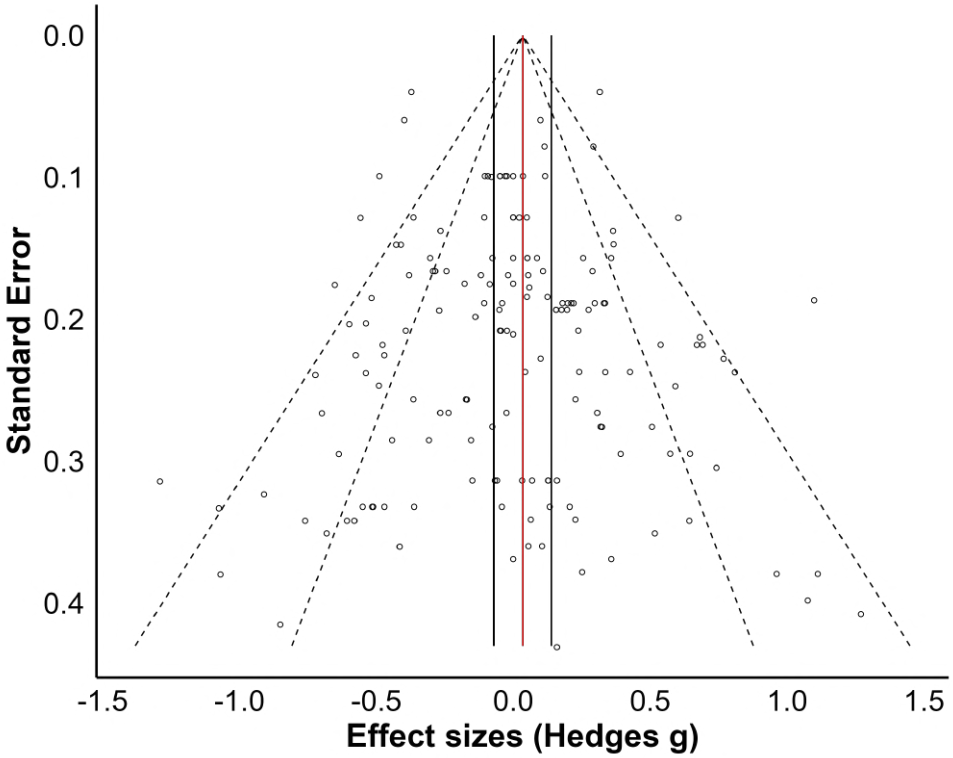
The included studies had a cumulative sample size of 14,313 participants. The average sample size per paper was approximately 117 (SD = 338.39), ranging from a minimum of 17 to a maximum of 2,445 participants. The average percentage of female participants was 58.96% (SD = 20.04), with a range from 0% to 96%. Gender distribution data were unavailable for 23 papers. The average reported age of participants was 21.39 years (SD = 10.51), with ages ranging from 8.2 to 64.4 years. Age data were not reported in 27 papers. The majority of the studies (76.5%) focused on university students. Detailed study characteristics that can serve as potential moderators are presented in the Moderators section.

3.2 Outlier detection and publication bias

The results of the outlier diagnostics, as illustrated in Figure 4.2a, indicate that no effect sizes deviated by more than three standard deviations from the mean group effect size of -0.009. Consequently, no effect sizes were removed at this stage.



(a) Outlier Check Histogram



(b) Funnel Plot

Figure 4.2: (a) Outliers check histogram and (b) Funnel plot.

To evaluate potential publication bias, we initially visualized the distribution of effect sizes against standard errors via a funnel plot, which appeared largely symmetrical (Figure 4.2b). The results of Egger's regression test revealed an intercept of 0.18 (SE=0.32, t=0.55, p=0.582). This suggests no significant deviation of the expected effect sizes from zero when precision is low, implying no disproportionate representation of small studies with large effects. Furthermore, the regression coefficient for precision (inverse SE) was -0.05 (SE=0.05, t=-0.92, p=0.360). This indicates that the effect sizes were not significantly influenced by the precision of the studies. These results suggest that there is no significant evidence of publication bias in this meta-analysis, and therefore, no adjustments for publication bias were necessary.

3.3 Overall Effect of Immediate versus Delayed Feedback on Learning (RQ1)

Using the Robust Variance Estimation (RVE) method, we calculated the overall weighted mean effect size from the 160 individual effect size estimates (from 51 studies), yielding $g = 0.03$ (95% CI [-0.07, 0.14], p=0.518) with an estimated between-study SE of 0.05. Heterogeneity was substantial ($I^2 = 82\%$) and the intra-study variance was notably high ($\tau^2 = 0.12$), suggesting that study conditions or populations may influence the effect of feedback timing, and inviting a moderator analysis.

Modifying the assumed within-study effect size correlation (ρ) did not affect the estimated value of g (see Appendix Table 7).

The forest plot displaying the weights and effect sizes of the included studies is provided in the Appendix 0.3.

3.4 Moderators

3.4.1 Moderating Effect of definitions of "immediate" and "delayed" feedback (RQ2)

Our reading of the literature has identified two ways in which the definition of feedback delay varied: the first was the unit in which the delay was defined (absolute time, number of intervening items, or a combination of both), and the second was the delay difference (between delayed and immediate feedback, which is itself sometimes delivered after a non-null delay).

Comparing delays defined in seconds (num. effects k=66) to those defined in items (k=92, reference category) revealed no significant difference ($g = -0.09$, $p = 0.32075$). This suggests that the unit in which the delay was defined, whether seconds or items, did not significantly impact the effect of feedback timing on learning outcomes. However, there was a significant effect of defining delay using both time and items compared to defining it solely by items ($g = -0.180$, $p = 0.0024$). Studies that defined delay using both time and items showed improved performance in conditions with feedback delay ($g = -0.15$, $p < 0.001$), whereas studies defining delay solely by either time ($g = -0.06$, $p = 0.421$) or items ($g = 0.030$, $p = 0.5702$) did not. However, this finding should be interpreted with caution due to the small sample size (k=2) of the mixed(item+seconds) group.

Analyses of the moderating effect of delay difference were first carried out separately for studies defining feedback delay by time and those defining it by items. Since delay differences

measured in time covered several orders of magnitude (from seconds to days) and tended to be highly skewed, we applied a log transformation to normalize their distribution and reduce the influence of extreme values. For studies defining delay by time (k=62 where the delay difference was reported), there was no significant effect of delay difference ($g = -0.010$, $p = 0.699$) (see section Delay Difference (Seconds(log)) in Table 4.2).

For the studies defining delay in items (k=80 where the delay difference was reported), there was a significant effect of delay difference ($g = -0.010$, $p = 0.0258$) (see section Delay Difference (Items) in Table 4.2). This means that the effect of delayed feedback on performance, although negative for small delay differences ($g = 0.19$ for 2 items of delay, meaning better performance for immediate feedback), increased by 0.01 standard deviations by item of delay. Thus, when the delay difference between immediate and delayed feedback increased beyond 21 items, it started having positive effects on performance (see Figure 4c).

In order to test the effect of delay difference across a maximum of studies irrespective of their definition of feedback delay, we attempted to convert item-based delays into equivalent times, for studies providing enough information on item timing to do so. This conversion was possible for 54 effects. For this subset of studies, where delays were converted from items to seconds and subsequently log-transformed, the moderating effect of delay difference was no longer significant ($g = -0.030$, $p = 0.603$), (see section Delay Difference (Items Converted to Seconds(log)) in Table 4.2).

Finally, when considering all studies expressing feedback delay in time or converted from items to seconds, totaling 116 effects (62 originally reported in seconds and 54 converted from items), there was no significant moderating effect of delay difference on learning outcomes ($g = -0.010$, $p = 0.747$), (see section Delay Difference (All Converted to Seconds(log)) in Table 4.2).

In conclusion, we found evidence in favor of the effect of delayed feedback on performance only in studies defining delay in items, and using a large number of intervening items.

Table 4.2: Preregistered Moderator Effects on Feedback Timing

Moderator	Estimate	Std. Error	t-value	$P(\|t\| >)$	95% CI	Num. Studies N	Num. Effects k
Delay Unit						51	160
Intercept(Item)	0.030	0.060	0.570	0.5702	[-0.08, 0.14]		92
Seconds (vs item)	-0.090	0.090	-1.000	0.3208	[-0.28, 0.09]		66
Mixed (item+seconds) (vs item)	-0.180	0.060	-3.210	0.0024**	[-0.29, -0.07]		2
Delay Difference (Seconds(log))						18	62
Intercept (Delay difference=0 sec)	-0.010	0.14	-0.070	0.948	[-0.3, 0.29]		
Delay difference	-0.010	0.02	-0.390	0.699	[-0.05, 0.04]		
Delay Difference (Items)						29	80
Intercept (Delay difference= 2 items)	0.19	0.09	2.09	0.0465*	[0, 0.38]		
Delay difference	-0.010	0.00	-2.360	0.0258*	[-0.02, 0]		
Delay Difference (Items Converted to Seconds(log))						16	54
Intercept (Delay difference=0 sec)	0.260	0.45	0.570	0.578	[-0.72, 1.23]		
Delay difference	-0.030	0.06	-0.530	0.603	[-0.17, 0.1]		
Delay Difference (All Converted to Seconds(log))						16	54
Intercept (Delay difference=0 sec)	0.020	0.140	0.130	0.899	[-0.27, 0.3]		
Delay difference	-0.010	0.020	-0.330	0.747	[-0.05, 0.04]		
Feedback Type						51	160
Intercept (Elaborated)	0.020	0.070	0.250	0.801	[-0.12, 0.16]		40
Try again (vs elaborated)	-0.410	0.070	-6.010	< 0.001**	[-0.55, -0.28]		2
Simple (vs elaborated)	-0.030	0.090	-0.350	0.727	[-0.21, 0.15]		118
Educational Level						51	160
Intercept (Tertiary)	0.030	0.050	0.600	0.5544	[-0.07, 0.13]		127
Secondary (vs tertiary)	-0.260	0.110	-2.270	0.0278*	[-0.49, -0.03]		22
Adult education (vs tertiary)	-0.120	0.060	-1.930	0.0597	[-0.24, 0.01]		6
Primary (vs tertiary)	-0.240	0.320	-0.750	0.4595	[-0.89, 0.41]		5
Educational Level (Regrouped)						51	160
Intercept (Tertiary+Adult)	0.022	0.045	0.484	0.6308	[-0.07, 0.11]		133
Primary+secondary (vs tertiary+adult)	-0.247	0.112	-2.206	0.0321*	[-0.47, -0.02]		27

Continued on next page

Continued from previous page

Moderator	Estimate	Std. Error	t-value	$P(\|t\| >)$	95% CI	Num. Studies N	Num. Effects k
Learning Domain						51	160
Intercept (Text memory)	-0.300	0.060	-5.030	< 0.001**	[-0.41, -0.18]		18
STEM (vs text memory)	0.290	0.100	2.850	0.00657**	[0.09, 0.5]		35
Social science (vs text memory)	0.140	0.060	2.430	0.01908*	[0.02, 0.26]		3
Reading comprehension (vs text memory)	-0.570	0.060	-9.740	< 0.001**	[-0.69, -0.45]		2
Language learning (vs text memory)	0.420	0.100	4.020	< 0.001**	[0.21, 0.63]		53
General cognitive skills (vs text memory)	0.250	0.090	2.660	0.01076*	[0.06, 0.44]		49
Learning Domain (Regrouped)						51	160
Intercept (Others)	0.030	0.050	0.550	0.582	[-0.07, 0.12]		20
Text-based learning (vs others)	-0.370	0.090	-4.310	< 0.001**	[-0.54, -0.20]		140
Training Task Complexity						51	160
Intercept (Higher order)	0.060	0.060	0.960	0.339	[-0.07, 0.19]		76
Lower order (vs higher order)	-0.130	0.090	-1.450	0.152	[-0.31, 0.05]		84
Prior Knowledge Adjustment						51	160
Intercept (Adjusted)	0.100	0.140	0.730	0.472	[-0.17, 0.37]		25
Not adjusted (vs adjusted)	-0.120	0.140	-0.860	0.393	[-0.41, 0.17]		135
Learning Task Context						51	160
Intercept (Curriculum based)	0.040	0.060	0.560	0.576	[-0.09, 0.16]		58
Experimental (vs Curriculum based)	-0.070	0.090	-0.830	0.409	[-0.24, 0.1]		102

Signif. codes: < .01 ** < .05 *

Table 4.3: Exploratory Moderator Effects on Feedback Timing

Moderator	Estimate	Std. Error	t-value	$P(\|t\| >)$	95% CI	Num. Studies N	Num. Effects k
Participant Design						51	160
Intercept (Between)	0.050	0.060	0.710	0.478	[-0.08, 0.17]		94
Within (vs between)	-0.110	0.090	-1.300	0.201	[-0.29, 0.06]		66
Learning Task Type						51	160
Intercept (Arbitrary)	0.030	0.100	0.320	0.750	[-0.17, 0.23]		30
Non-arbitrary (vs arbitrary)	-0.050	0.110	-0.480	0.633	[-0.28, 0.17]		130
Post-test Item Similarity						47	154
Intercept (Different from training)	0.080	0.070	1.060	0.296	[-0.07, 0.23]		55
Identical to training (vs different from training)	-0.140	0.100	-1.410	0.165	[-0.33, 0.06]		80
Mixed (vs different from training)	-0.110	0.180	-0.620	0.541	[-0.47, 0.25]		19
Feedback Dependency on Errors						51	160
Intercept (Independent of error)	-0.020	0.050	-0.360	0.720	[-0.11, 0.08]		126
Only for errors (vs independent of error)	0.040	0.140	0.300	0.762	[-0.24, 0.33]		34
Feedback Answer Reminder						35	109
Intercept (No answer reminder)	-0.040	0.080	-0.530	0.598	[-0.22, 0.13]		65
Answer reminder (vs no answer reminder)	0.020	0.110	0.170	0.870	[-0.21, 0.25]		44
Nb Training Items						46	148
Intercept (Nb training items=4)	-0.010	0.060	-0.090	0.931	[-0.13, 0.12]		
Nb training items	-0.000	0.000	-0.240	0.814	[0, 0]		
Post-test Task						51	160
Intercept (Knowledge Application)	0.090	0.060	1.500	0.1398	[-0.03, 0.21]		74
Memory retrieval (vs knowledge application)	-0.180	0.080	-2.120	0.0389*	[-0.35, -0.01]		86
Time Limitation for Responses						51	160
Intercept (Limited)	0.040	0.070	0.640	0.5242	[-0.09, 0.18]		60
Not limited (vs limited)	-0.190	0.080	-2.270	0.0276*	[-0.36, -0.02]		42
Not reported (vs limited)	0.000	0.110	0.030	0.9732	[-0.22, 0.23]		58

Continued on next page

Continued from previous page

Moderator	Estimate	Std. Error	t-value	$P(\|t\| >)$	95% CI	Num. Studies N	Num. Effects k
Retention Interval Difference btw Last Feedback and Post-test(log))						33	110
Intercept (Retention interval difference= 0 sec)	-0.020	0.050	-0.420	0.681	[-0.13, 0.09]		
Retention interval difference	-0.010	0.010	-0.840	0.409	[-0.02, 0.01]		
Publication Year						51	160
Intercept (Publication Year =1995)	-0.36	0.10	-3.500	0.00100**	[-0.57, -0.15]		
Publication Year	0.020	0.010	3.400	0.00135**	[0.01, 0.03]		

Signif. codes: < .01 ** < .05 *

3.4.2 Other Preregistered Moderators (RQ3)

In addressing our third research question, we first evaluated the impact of preregistered moderators on the comparative effectiveness of immediate versus delayed feedback on learning outcomes. The distributions of effect sizes for the eight preregistered moderators are comprehensively detailed in the Appendix, as illustrated in Figure 4.

Among the examined moderators, a significant effect was found when comparing try-again feedback (k= 2) to elaborated feedback (k=40, reference category), ($g = -0.410$, $p < 0.001$). Studies employing try-again feedback demonstrated improved performance under conditions with delayed feedback ($g = -0.40$, $p < 0.001$), whereas studies using elaborated feedback did not ($g = 0.02$, $p = 0.801$, see Table 4.2). However, caution is warranted in interpreting these results due to the limited sample size for the try-again feedback type. Furthermore, when comparing simple feedback (k= 118) to elaborated feedback, the difference was not statistically significant ($g = -0.030$, $p = 0.727$). This suggests that the type of feedback used in the experiment, whether simple or elaborated, does not significantly influence the effect of feedback timing on learning outcomes.

Another significant moderating effect was observed when comparing studies conducted at the secondary (k= 22) with those at the tertiary education levels (k= 127, reference category), ($g = -0.260$, $p = 0.0278$). This suggests that studies involving secondary education participants demonstrated greater improvements in performance under delayed feedback conditions ($g = -0.23$, $p = 0.0305$), whereas studies with tertiary education participants showed no significant advantage for either immediate or delayed feedback ($g = 0.030$, $p = 0.5544$). Additionally, again using studies at the tertiary education level as the reference category, studies conducted in adult education (k= 6) and primary education (k= 5) showed a nominal but non-significant advantage for delayed feedback ($g = -0.120$, $p = 0.0597$) and ($g = -0.240$, $p = 0.4595$), respectively, based on relatively few numbers of effects.

For the purpose of the subsequent meta-regression, we grouped together similar education levels into just 2 categories: 1) primary + secondary and 2) tertiary + adult. This grouped version of the moderator was re-tested and revealed that delayed feedback significantly improved learning outcomes in studies involving participants in primary or secondary education compared to those in tertiary or adult education ($g = -0.247$, $p = 0.0321$) (see section Educational Level (grouped) in Table 4.2).

The moderator analysis revealed a differential impact of feedback timing across learning domains. Compared to experiments in which the domain was text memorization (k= 18, reference category), studies on general cognitive skills (k= 49), ($g = 0.250$, $p = 0.01076$), language learning (k=53), ($g = 0.420$, $p < 0.001$), social sciences (k= 3), ($g = 0.140$, $p = 0.01908$), and STEM (k= 35), ($g = 0.290$, $p = 0.00657$) demonstrated a significant advantage of immediate feedback. Conversely, compared to text memorization studies, studies on reading comprehension (k= 2) exhibited an even greater and significant advantage for delayed feedback ($g = -0.570$, $p < 0.001$). However, examining the absolute effect of feedback timing per learning domain, irrespective of the comparison, revealed that reading comprehension ($g = -0.87$, $p < 0.001$), social science ($g = -0.15$, $p < 0.001$), and text memorization ($g = -0.30$, $p < 0.001$) all demon-

strated a significant advantage for delayed feedback. Conversely, cognitive skills ($g = -0.05$, $p = 0.5420$), language learning ($g = 0.13$, $p = 0.15196$), and STEM ($g = 0.00$, $p = 0.99809$) did not exhibit significant advantages in feedback timing.

For the subsequent meta-regression analysis, the learning domain moderator was restructured by combining reading comprehension and text memorization into a single category labeled "text-based learning", while all other domains were grouped under "others". This revised moderator analysis revealed that, in experiments focusing on text-based learning, delayed feedback had a significant advantage compared to other domains ($g = -0.370$, $p < 0.001$) (see Section Learning Domain (grouped) in Table 4.2).

Regarding training task complexity, there was no significant difference in feedback delay effectiveness between lower-order tasks (k=84) and higher-order tasks (k=76), ($g = -0.160$, $p = 0.333$). Similarly, prior knowledge adjustment did not significantly moderate the effectiveness of feedback timing, as shown by the comparison of unadjusted (k=135) to adjusted outcomes (k=25), ($g = -0.120$, $p = 0.393$). Finally, learning task context did not moderate the impact of feedback timing, with no significant difference observed between experimental tasks (k=102) and curriculum-based tasks (k=58), ($g = 0.070$, $p = 0.576$).

3.4.3 Exploratory Moderators (RQ3)

Beyond the preregistered moderators, we examined ten exploratory moderators identified through the screening procedure. As shown in Table 4.3, only three were found to significantly moderate the effect of feedback timing on learning outcomes. Effect size distributions for all these exploratory moderators are provided in the Appendix (Figure 5)

Post-test task type emerged as a significant moderator, with improved performance in the conditions with feedback delay in studies with memory retrieval post-test tasks (k= 86) than in those with knowledge application post-test tasks (k= 74, reference category), ($g = -0.180$, $p = 0.0389$). However, when analyzing the absolute effect of feedback timing on learning outcomes within each post-test task type, no significant effects were found. In memory retrieval tasks, delayed feedback showed a slight but non-significant advantage ($g = -0.09$, $p = 0.1544$), while in knowledge application tasks, immediate feedback appeared more effective ($g = 0.090$, $p = 0.1398$), though this effect did not reach statistical significance.

Secondly, a significant effect was observed when comparing studies without a time limit for responses (k=42) to those with a limited response time (k=60, reference category), ($g = -0.190$, $p = 0.0276$). Studies that did not impose a time limit for responses demonstrated improved performance under delayed feedback conditions ($g = -0.15$, $p = 0.0051$), whereas studies with a limited response time showed no significant advantage for either immediate or delayed feedback ($g = 0.040$, $p = 0.5242$).

Finally, the publication year of the studies emerged as a significant moderator ($g = 0.020$, $p = 0.00135$), indicating that the effect of delayed feedback on performance, though positive in older studies (intercept at the minimum publication year, 1995, with $g = -0.36$, suggesting better performance under delayed feedback), decreased by 0.02 standard deviations per year. It reached approximately zero around 2013, after which more recent studies found no benefit for

delayed feedback (see Figure 5j).

3.4.4 Multiple Meta-regression

In our multiple meta-regression, we included only the four statistically significant moderators: educational level, learning domain, post-test task, and time limitation for responses. The publication year moderator, although significant in the individual analyses, was excluded, as it is not an experimental or feedback-related property but rather a factor likely to covary with other moderators.

To facilitate the interpretation of the multiple meta-regression, we used the grouped versions of certain moderators (educational level and learning domain) to ensure that only binary moderators were included.

No grouping was performed for the post-test task or time limitation for response moderators. Notably, the "Not Reported" category in the time limitation for responses moderator was retained to account for the substantial number of effect sizes ($n = 58$) with missing data. This inclusion ensured that the statistical power of the multiple meta-regression was not reduced by excluding these cases.

While individual analyses identified significant moderating effects for these four variables, the results of the multiple meta-regression (Table 4.4) indicate that none of the moderators remained statistically significant. This suggests that these moderators are partially confounded with one another.

To further explore the potential confounding relationships among the moderators included in the multiple meta-regression, we first analyzed the distribution of educational level, learning domain, post-test task type, and time limitation for responses in relation to one another using a contingency table (Table 4.5). The table revealed notable dependencies between certain moderators. Specifically, educational level and learning domain were closely related ($\chi^2(1) = 41.76, p < .001$), as studies in tertiary/adult education primarily focused on non-text-based learning (k=127), whereas those in primary/secondary education were more evenly distributed across learning domains. Similarly, post-test task type and time limitation for responses were linked ($\chi^2(2) = 56.10, p < .001$), with knowledge application post-test tasks more frequently conducted with the training under time-limited conditions (k=22).

Additionally, the publication year distribution of each moderator was examined (Figure 4.3) to assess whether the publication year of the studies may have systematically influenced moderator effects. The distributions indicate a noticeable shift over time, with an increasing number of studies involving tertiary/adult participants ($t(30.81) = 4.65$, $p < .001$, $d = 1.28$), non-text-based learning tasks ($t(22.01) = -6.33$, $p < .001$, $d = -1.91$), post-test tasks requiring knowledge application ($t(151.95) = 5.28$, $p < .001$, $d = 0.82$), and responses in a limited time ($t(73.89) = 4.24$, $p < .001$, $d = 0.89$).

Table 4.4: Multiple Meta-Regression

Moderator	**Estimate**	**Std. Error**	**t-value**	$P(\|t\| >)$	**95% CI**
Intercept	0.01	0.13	0.05	0.959	[-0.25, 0.27]
Educational Level (Regrouped)					
Tertiary+adult (vs primary+secondary)	0.15	0.14	1.05	0.297	[-0.14, 0.44]
Learning Domain (Regrouped)					
Text-based learning (vs others)	-0.20	0.16	-1.23	0.227	[-0.52, 0.13]
Post-test Task					
Memory retrieval (vs knowledge application)	-0.15	0.10	-1.41	0.164	[-0.36, 0.06]
Time Limitation for Responses					
Not limited (vs limited)	-0.10	0.09	-1.19	0.242	[-0.28, 0.07]
Not reported (vs limited)	-0.05	0.11	-0.47	0.644	[-0.28, 0.18]
Num. Clusters	51				
Num. Obs.	160				
Signif. codes:	$< .01$ ** $< .05$ *				

Table 4.5: Contingency Table for Multiple Meta-regression Moderators with Chi-Square Results (Alternative)

	Learning Domain		**Post-test Task**		**Time Limitation for Responses**		
	Others	Text-based	Application	Retrieval	Limited	Not Limited	Not Reported
Educational Level	$\chi^2(1) = 41.76, p < .001$		$\chi^2(1) = 0.00, p = .996$		$\chi^2(2) = 10.40, p = .006$		
Tertiary + Adult	127	6	61	72	56	29	48
Primary + Secondary	13	14	13	14	4	13	10
Learning Domain			$\chi^2(1) = 10.47, p = .001$		$\chi^2(2) = 17.79, p < .001$		
Others	–	–	72	68	56	29	55
Text-based Learning	–	–	2	18	4	13	3
Post-test Task					$\chi^2(2) = 56.10, p < .001$		
Knowledge Application	–	–	–	–	22	4	48
Memory Retrieval	–	–	–	–	38	38	10

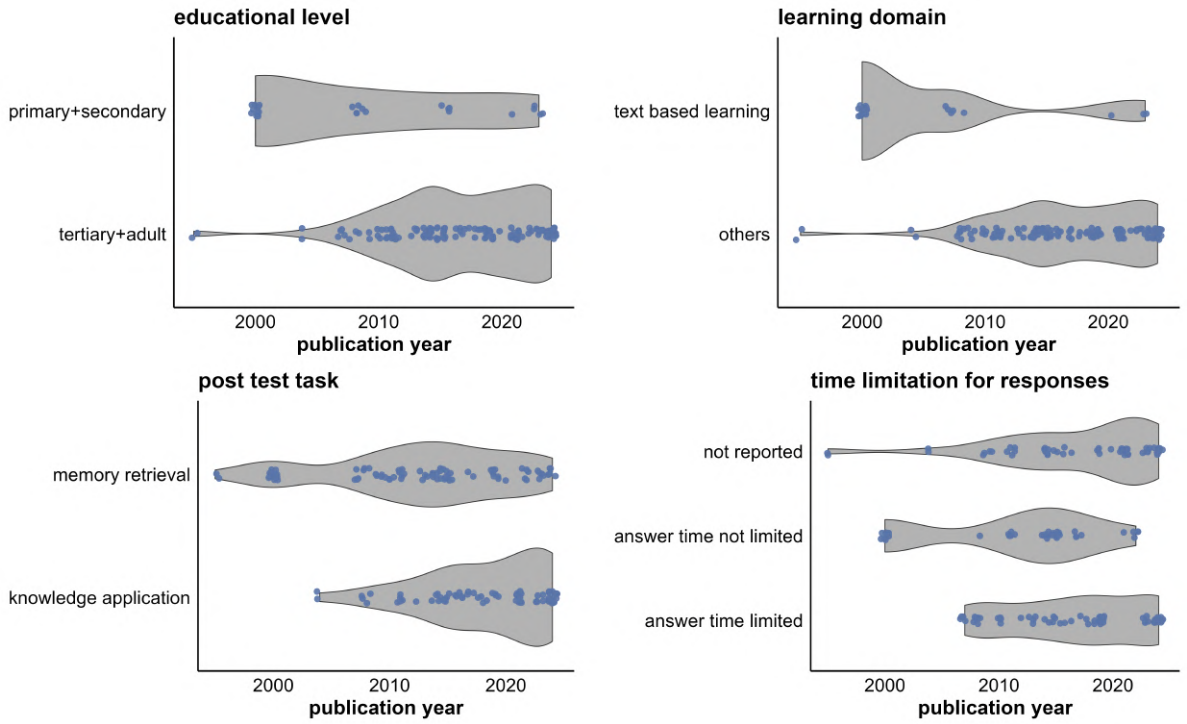


Figure 4.3: Distribution of Multiple Meta-regression Moderators across Publication Years

4 Discussion

This meta-analysis systematically examined the impact of immediate versus delayed feedback on learning outcomes. The results indicate that feedback timing does not consistently affect learning, as evidenced by the absence of a significant overall effect and substantial heterogeneity among studies. However, factors such as educational level, learning domain, post-test task type, and response time constraints significantly moderated the effect, partly explaining the observed variability.

This study builds on previous research on feedback timing by systematically examining a broader range of potential moderating factors. One key advancement is its consideration of the variability in how studies define immediate and delayed feedback to address a significant gap in prior meta-analyses and inconsistencies in the literature. Furthermore, this meta-analysis explored various moderators to explain the observed heterogeneity, including feedback characteristics (e.g., delay unit, feedback type) and study features (e.g., educational level, learning domain, post-test task type, and response time constraints).

4.1 Moderator effects in univariate analyses

Based on prior research, we had preregistered and systematically coded a number of moderators hypothesized to influence feedback timing effects. One such factor was the definition of immediate and delayed feedback. The literature highlights inconsistencies in how these terms are operationalized, with variations in delay units (items or time) and delay length. As a result,

what is considered immediate feedback in some studies may be classified as delayed feedback in others. This inconsistency has been cited as a primary reason for mixed findings in feedback timing research (Mory, 2013; M. Xu et al., 2023; Nakata, 2015). Quinn (2014) further argued that many studies do not introduce a sufficient temporal gap to establish a meaningful cognitive distinction between immediate and delayed feedback. Similarly, Canals et al. (2021) found no significant differences in learning outcomes, likely due to unsufficiant delay differences between conditions. To clarify this point, we categorized immediate and delayed feedback based on delay unit (items vs. time) and on the delay difference between feedback conditions. Our moderator analysis found no overall difference in outcomes based on whether delay was defined by items or time. However, the relative difference between immediate and delayed feedback—specifically in terms of the number of intervening items—significantly moderated learning outcomes, but only in studies that used item-based delay definitions. In contrast, when delay was defined in seconds or when item-based delays were converted into estimated time durations, this effect was no longer observed. These findings underscore the significant variability in how studies define and measure feedback delay, particularly when comparing item-based and time-based metrics. This inconsistency complicates the interpretation and comparison of research findings in feedback timing. Notably, the lack of a reliable correlation between the number of intervening items and the actual time elapsed before feedback delivery suggests that item-based delays can vary widely in duration across different studies. For instance, one study might implement a delay of several items that correspond to a few minutes, while another similar item-based delay could span a longer time due to differences in task complexity or participant response rates. This discrepancy indicates that item-based measures alone may not accurately reflect the temporal aspects of feedback delays. Therefore, to facilitate meaningful comparisons and syntheses across studies, it is advisable to accompany item-based delay measures with precise indications of the corresponding time durations.

An important characteristic predicted to moderate the effects of feedback timing—based on prior research was feedback type. However, we found no evidence that the type of feedback (simple vs. elaborated) moderated the impact of feedback timing on learning outcomes. While previous meta-analyses have shown that elaborated feedback is generally more effective than simple feedback (Brummer et al., 2024; M. Xu et al., 2023; Cai et al., 2023; Wisniewski et al., 2020), these studies focused on the main effects of feedback type rather than its interaction with timing. One exception is Taxipulati et al. (2021), who reported that learners responded more attentively to adaptive feedback when it was provided immediately rather than after a delay—suggesting a possible interaction between timing and feedback type under specific conditions. However, they did not report whether the timing of feedback (immediate vs. delayed) differentially impacted learning outcomes when comparing elaborated feedback to knowledge-of-correct-response feedback. Furthermore, our meta-analysis did not categorize adaptive feedback separately, as this feature was rarely reported across the included studies. The inconsistent reporting of adaptive feedback may have limited our ability to detect such interactions. Nonetheless, our findings are not necessarily at odds with the existing literature.

Educational level—another factor hypothesized to moderate the effects of feedback tim-

ing—emerged as a significant moderator in our analysis. Specifically, delayed feedback was associated with significantly better learning outcomes among secondary education students, whereas no timing effect was observed in tertiary-level learners. These findings suggest that learners at different educational stages may process and benefit from feedback timing in distinct ways, possibly due to developmental differences in cognitive maturity, self-regulation, and engagement. This pattern aligns with prior claims that the impact of feedback may vary across educational levels (Swart et al., 2019). While Kulik et al. (1988) previously found no moderating effect of educational level, our findings may reflect shifts in educational practices and learner profiles since then—particularly in light of the increased integration of technology in instruction. Moreover, this pattern may partly reflect a temporal shift in participant profiles in more recent studies that were not included in Kulik's meta-analysis. This covariance between educational level and publication year is discussed further in Section 4.2.

As outlined in the introduction, we also hypothesized that the optimal timing of feedback would vary across learning domains due to differences in task characteristics and cognitive demands (Van der Kleij, Feskens, et al., 2015; Mertens et al., 2022). Consistent with this hypothesis, our analysis found that the learning domain significantly moderated the effect of feedback timing. Delayed feedback was more beneficial for tasks involving reading comprehension, social sciences, and text memorization. In contrast, domains such as cognitive skills, language learning, and STEM did not show a consistent advantage for either immediate or delayed feedback. These domain-specific effects may reflect differences in how learning unfolds across subject areas. In domains like reading comprehension and social sciences—where tasks often require meaning-making, reflection, and integration of complex information—delayed feedback may promote deeper processing by allowing time for retrieval and self-monitoring, consistent with theories of spacing and desirable difficulties (E. L. Bjork et al., 2011; Carpenter, 2014). However, the small number of comparisons in some domains (e.g., reading comprehension and social sciences) limits the generalizability of these findings and warrants cautious interpretation. In contrast, domains like STEM and language learning, which were better represented in the dataset, may involve more procedural or fluency-based tasks where feedback timing plays a less consistent role. The absence of a clear timing effect in these areas may also reflect greater heterogeneity in task types and instructional designs.

Training task complexity, categorized as higher or lower order according to Bloom et al. (1964), was another preregistered moderator that showed no significant effect. Based on prior research, we expected a moderating pattern, similar to that observed for post-test task types. Specifically, we predicted that immediate feedback would be more effective for lower-order tasks, whereas delayed feedback would yield greater benefits for higher-order tasks (Shute, 2008; Van der Kleij, Feskens, et al., 2015). However, our results did not support this expectation. One possible explanation is that our analysis focused on post-test outcomes, while training task complexity may be more relevant during the acquisition phase, warranting further investigation.

Adjustment of effect sizes on prior knowledge was also preregistered as a potential moderator, given the suggested interaction between feedback timing and learner ability (Nathan et al., 2002). However, our findings showed no significant moderating effect, suggesting that the experimental

studies included in our meta-analysis effectively randomized participants across feedback timing conditions, as expected.

Lastly, we explored whether the nature of the learning task —whether designed purely for research purposes or integrated into a real curriculum — acted as a moderator. This expectation was based on Kulik et al. (1988), which found that applied settings generally favored immediate feedback, whereas laboratory studies tended to support delayed feedback. We hypothesized that tasks relevant to real-world applications and directly beneficial to learners' educational or professional goals might influence motivation and cognitive load, potentially affecting the efficacy of immediate versus delayed feedback differently. However, our analyses did not support this hypothesis.

Although not preregistered, we explored several additional moderators that appeared to vary systematically across studies during the screening process and that have theoretical grounding and prior support in the literature as potential explanations for inconsistencies in feedback timing effects.

One such moderator was the type of the post-test task, which emerged as a significant moderator in our analysis. When the post-test required applying knowledge, immediate feedback was more beneficial than when it focused solely on recall. However, neither immediate nor delayed feedback consistently outperformed the other when examined within each task type individually. This finding contrasts with Shute (2008), who argued that delayed feedback better promotes knowledge transfer. Our results highlight the complex interplay between feedback timing and task demands, underscoring the need for further investigation.

Another significant moderator was the presence or absence of response time constraints during training. We observed that studies allowing unlimited response time tended to report stronger effects for delayed feedback compared to studies with time-limited responses. Although this variable has not been previously examined in meta-analyses on feedback timing, our findings align with cognitive load theory (Sweller, 1988; Sweller et al., 2019). Without time constraints, learners can engage in deeper cognitive processes, such as self-questioning and integrating new information, potentially enhancing the benefits of delayed feedback. In contrast, limited response time may increase cognitive load and hinder deep processing, reducing the effectiveness of delayed feedback. Notably, around 30% of studies did not specify response time constraints, indicating a gap in the research. Future studies should explicitly report this variable, given its significant influence on feedback effectiveness.

We also examined publication year as a potential moderator and found evidence that the effect of feedback timing has shifted over time. Earlier studies tended to report stronger benefits of delayed feedback. In contrast, more recent studies have increasingly favored immediate feedback or found no timing advantage, potentially due to the growing integration of digital technologies that enable real-time interaction, as well as shifts in learner expectations and habits. This temporal trend suggests that the effectiveness of feedback timing may be context-dependent and influenced by broader changes in educational environments and research methodologies. To further explore this possibility, we examined whether publication year correlated with other significant moderators—this is reported below in multivariate analyses of moderators.

In addition to the significant moderators above, we also explored several theoretically grounded variables that did not show significant moderating effects but warrant discussion.

One such variable was the presence of an error reminder. Our reading of the literature suggested that, in studies where immediate feedback was more effective, delayed feedback was often provided by reminding the learner's error before presenting the correction (e.g., (Li et al., 2016)). In contrast, studies using delayed feedback often did not prompt learners to recall their initial answers before or during feedback (e.g., (Carpenter and Vul, 2011; A. C. Butler, Karpicke, et al., 2007)). This pattern aligned with the interference-perseveration hypothesis (Kulhavy and R. C. Anderson, 1972), which suggests that error reminders may enhance the effects of immediate feedback, whereas their absence could favor delayed feedback. However, our meta-analysis found no significant moderating effect of the presence of an error reminder.

We also examined the retention interval difference between feedback and post-test as a potential moderator, given that shorter delays generally enhance memory retention (Nicholas J. Cepeda et al., 2008; Metcalfe, Kornell, et al., 2009). Some studies suggest that controlling for lag-to-test intervals reduces differences in effectiveness between immediate and delayed feedback (Metcalfe, Kornell, et al., 2009; Nakata, 2015). However, many studies confound delayed feedback with extended retention intervals, making it unclear whether observed benefits stem from feedback timing or time elapsed before testing. Similarly, training duration may influence feedback effects, as longer training phases in delayed conditions inherently extend retention intervals, potentially impairing recall (Nicholas J. Cepeda et al., 2008; Rohrer et al., 2005). Although our meta-analysis did not find significant moderating effects for either retention interval or training length, these null results may reflect limited statistical power due to a small number of studies that reported or controlled for these variables. Future research should report these factors more systematically to enable a clearer interpretation of their role in feedback effectiveness.

4.2 Moderator effects in multivariate analyses

While individual moderators like educational level, learning domain, post-test task type, and response time constraints showed significant moderating effects when analyzed separately, their respective influence decreased in the multiple meta-regression. This discrepancy, coupled with the significant moderating effect of the publication year, led us to investigate potential interdependencies between these study characteristics. Indeed, our analysis revealed that these factors were not independent from each other, but rather exhibit significant covariation. For example, educational level and learning domain were closely associated, with tertiary/adult education studies primarily focusing on non-text-based learning, while primary/secondary studies were more varied. Similarly, post-test task type and response time constraints were related, as knowledge application tasks were more often performed under time-limited conditions.

Furthermore, an analysis of publication trends revealed systematic shifts in study characteristics over time. Experimental paradigms shifted from examining text-based learning and assessing outcomes through memory retrieval tasks without time constraints in primary/secondary school pupils to studies examining other types of learning domains, testing knowledge application under time-limited conditions, in tertiary/adult participants.

These dependencies suggest that it is not possible to completely disentangle the effects of these various moderators based on the existing set of studies. Rather, it is possible to characterize the studies that typically yield greater delayed vs. immediate feedback effects.

Studies that suggested an advantage of delayed feedback were typically published before 2013, involved text memorization or reading comprehension, memory retrieval rather than knowledge application, no response time limit, defined feedback delay in number of items, defined delayed feedback by a large (> 21) number of items, and were carried out on primary/secondary pupils. On the contrary, studies that suggest no difference between delayed and immediate feedback tended to be published more recently, involved learning in cognitive skills, second language or STEM, involved knowledge application tasks and were carried out on tertiary/adult participants. Finally, studies that suggested an advantage of immediate feedback defined feedback delays in terms of a small number of items.

The overall findings of this meta-analysis partially align with previous research on feedback timing, but also help reconcile their inconsistencies by identifying patterns across study characteristics. Earlier meta-analyses, such as Azevedo et al. (1995), reported a clear advantage for immediate feedback, while others, like Swart et al. (2019), found benefits for delayed feedback in specific domains such as reading comprehension. More recently, reviews like M. Xu et al. (2023) have concluded that immediate feedback is often more or equally effective, depending on contextual factors.

Our results suggest that these differing conclusions reflect systematic variation in study-level characteristics rather than a universal advantage of one feedback timing over another. These patterns suggest that the effectiveness of feedback timing is highly context-dependent and has evolved alongside shifts in learning context, participant populations, and instructional design within computer-based learning environments over time.

4.3 Limitations and Directions for Future Research

Our study is subject to several limitations. One primary limitation is the substantial heterogeneity observed across studies, which was not adequately accounted for by most moderator analyses. A key contributing factor was the inconsistent reporting of critical study variables—particularly among categorical moderators—which constrained our ability to conduct more comprehensive meta-regressions. Additionally, some studies lacked sufficient data to compute effect sizes despite meeting other inclusion criteria. As a result, the small number of effect sizes reduced the reliability of certain findings, making it difficult to determine whether observed effects were robust or a result of limited data. This issue was particularly evident in comparisons of feedback timing effects across educational levels. The majority of studies focused on university students, with relatively few examining primary education ($n = 5$), limiting the ability to draw conclusions about younger learners. Similarly, certain learning domains, such as social sciences ($n = 3$) and reading comprehension ($n = 2$), were underrepresented, reducing the generalizability of findings. With such limited power, we can only conclude that additional studies are needed to improve the reliability of the findings (Harrer et al., 2021).

Another limitation concerns the inconsistent conceptualization of feedback timing in the lit-

erature. To accommodate this variation, we categorized feedback delays into time-based and item-based groups. This division, however, reduced statistical power. To regain power, we attempted to approximate item-based delays in temporal terms using additional study information. Unfortunately, these details were not consistently reported, making such conversions unreliable. Thus, more experimental studies are needed to clarify inconsistencies in feedback delay definitions and their effects on learning outcomes since meta-analyses cannot generate new findings beyond what is available in the experimental literature.

A further limitation relates to the range of moderators examined. Prior research suggests that various feedback characteristics, study features, and learner attributes may influence the effectiveness of feedback, but many of these could not be systematically coded due to insufficient reporting. For instance, Clariana et al. (2000) suggested that learning material difficulty may influence optimal feedback timing. Yet, comparing difficulty levels across studies is inherently challenging, as they are often not measured on a common scale and are operationalized differently depending on the context. Furthermore, C. Timmers et al. (2011) proposed that learners' prior knowledge could moderate feedback effectiveness, while Nathan et al. (2002) highlighted the interaction between feedback timing and learner ability. However, as difficulty level and learner ability were not systematically reported, moderator analyses on these factors were not feasible. Future research should investigate these moderators in both primary studies and meta-analyses.

Furthermore, the findings of this meta-analysis are based primarily on controlled experimental settings, particularly in computer-based tutoring systems. Most studies measured effects in the short term, using posttests administered immediately or shortly after feedback interventions. This approach may not fully capture real-world educational complexities, where factors like classroom dynamics, student motivation, and instructor feedback styles play a role. Thus, the ecological validity of our results is limited.

Finally, this meta-analysis focused solely on feedback timing's impact on learning outcomes, as measured by posttest scores. Future research should also examine its effects on student engagement, as studies suggest that immediate feedback is often preferred by educators and students (Lefevre et al., 2017; Van der Kleij, Eggen, et al., 2012; Mullet et al., 2014). Investigating the balance between learning effectiveness and engagement would provide a more comprehensive understanding of feedback timing in educational contexts.

Given that our multivariate moderator analysis was limited by the fact that moderators were partly confounded with each other, providing more definitive conclusions about the respective role of each moderator would require new studies with specific characteristics that were missing in the current literature. In particular, it would be useful to have more studies comparing immediate and delayed feedback: 1) in primary and secondary school pupils in general; 2) involving text-based learning in tertiary/adult participants; 3) involving limited response time in primary/secondary school participants; 4) involving text-based learning and limited response time; 5) involving knowledge application tasks and unlimited response time; 6) involving text-based learning and knowledge application tasks.

4.4 Practical Implications

Despite the limitations, the findings of this meta-analysis have some implications for the design and implementation of educational technologies, particularly in computer-assisted learning environments where feedback timing can be precisely controlled to improve learning.

Although the present meta-analysis cannot provide complete recommendations concerning the specific conditions in which delayed or immediate feedback yield better learning in all possible conditions, it can at least provide partial indications for some types of studies. In particular, for learning contexts that involve text memorization or reading comprehension, memory retrieval outcomes, and no response time limit, it may be advantageous to use delayed feedback with a large (> 21) number of intervening items. On the contrary, in learning contexts where it is not feasible to delay feedback with a large number of intervening items, then sticking to immediate feedback may be preferable. For most other learning contexts, our meta-analysis does not provide strong reasons to prefer immediate or delayed feedback.

As research evolves, refining feedback strategies to align with modern educational practices remains essential. Developers should continue integrating evidence-based approaches to enhance feedback effectiveness in digital learning tools.

5 Appendix

0.1 Search terms

- **Web of Science:**

```
(TS=("feedback" OR "assessment" OR "correction") AND
TS=("feedback timing" OR "feedback time" OR "time of feedback" OR
"timing of feedback" OR (immediate AND delayed)) AND
PY=(1988-2025) AND
LA=(English OR Turkish OR French))
```

Results: 3,812 documents.

- **Scopus:**

```
(TITLE-ABS-KEY (feedback OR assessment OR correction) AND
TITLE-ABS-KEY ("feedback timing" OR "feedback time" OR
"time of feedback" OR "timing of feedback" OR (immediate AND delayed)) AND
PUBYEAR > 1987 AND
LANGUAGE (English OR Turkish OR French) AND
NOT SUBJAREA (MEDI OR VETE OR PHAR OR DENT OR NURS))
```

Results: 1,993 documents.

- **ERIC:**

```
((abstract:(feedback OR assessment OR correction)) AND
((abstract:(immediate AND delayed)) OR
(abstract:("feedback timing" OR "feedback time" OR
"time of feedback" OR "timing of feedback"))))
pubyearmin:1988
```

Results: 617 documents.

0.2 Included Studies

Table 6: Studies included in the meta-analysis

Study ID / Experiment	g	SE	Delay Unit	Delay Difference	Feedback Type	Educational Level	Learning Domain	Task Complexity	Prior Know. Adj.	Task Context
Albrecht et al. (2023)										
Experiment1	-0.18	0.18	sec.	6	SF	adult ed.	GCS	LO	no	Exp
Experiment1	00	0.18	sec.	6	SF	adult ed.	GCS	LO	no	Exp
Aljabri (2024)										
Experiment1	0.54	0.22	item	14	SF	tertiary	LL	LO	no	Exp
Experiment1	0.77	0.23	item	14	SF	tertiary	LL	LO	no	Exp
Experiment1	0.69	0.22	item	14	SF	tertiary	LL	LO	no	Exp
Arroyo et al. (2018)										
Experiment1	0.96	0.38	item	11	SF	tertiary	LL	HO	yes	CB
Experiment1	1.11	0.38	item	11	SF	tertiary	LL	HO	yes	CB
Experiment1	05	0.36	item	11	SF	tertiary	LL	HO	yes	CB
Experiment1	0.11	0.36	item	11	SF	tertiary	LL	HO	yes	CB
Attali and Kleij (2017)										
Experiment1	0.32	04	item	19	EF	adult ed.	STEM	HO	yes	Exp
Experiment1	-0.37	04	item	19	EF	adult ed.	STEM	HO	yes	Exp
A. C. Butler, Karpicke, et al. (2007)										
Experiment1	-0.15	0.29	sec.	600	SF	tertiary	TM	LO	no	Exp
Experiment1	-0.31	0.29	sec.	600	TAF	tertiary	TM	LO	no	Exp
Experiment2	-0.58	0.23	sec.	86400	SF	tertiary	TM	LO	no	Exp
Experiment2	-0.47	0.23	sec.	86400	TAF	tertiary	TM	LO	no	Exp
A. C. Butler and Roediger (2008)										
Experiment1	-0.56	0.13	item	12	SF	tertiary	TM	LO	no	Exp
Canals et al. (2021)										
Experiment1	0.52	0.35	sec.	86400	SF	tertiary	LL	HO	yes	CB
Experiment1	0.23	0.34	sec.	86400	SF	tertiary	LL	HO	yes	CB
Candel, Vidal-Abarca, et al. (2020)										
Experiment1	0.34	0.24	item	3	SF	tertiary	TM	HO	no	Exp
Candel, Máñez, et al. (2021)										
Experiment1	-0.39	0.21	item	3	EF	secondary	STEM	LO	no	Exp
Carpenter and Vul (2011)										
Experiment1	-0.52	0.19	sec.	3	SF	tertiary	GCS	LO	no	Exp
Experiment2	-0.43	0.15	sec.	3	SF	tertiary	GCS	LO	no	Exp

Table 6 continued from previous page

Study ID / Experiment	g	SE	Delay Unit	Delay Difference	Feedback Type	Educational Level	Learning Domain	Task Complexity	Prior Know. Adj.	Task Context
Experiment2	-0.27	0.14	sec.	3	SF	tertiary	GCS	LO	no	Exp
Experiment2	-0.41	0.15	sec.	3	SF	tertiary	GCS	LO	no	Exp
Experiment3	-0.29	0.17	sec.	3	SF	tertiary	GCS	LO	no	Exp
Experiment3	0.29	0.17	sec.	3	SF	tertiary	GCS	LO	no	Exp
Clariana et al. (2000)										
Experiment1	-0.76	0.34	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-0.47	0.33	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-0.55	0.33	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-0.61	0.34	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-0.58	0.34	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-0.52	0.33	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-04	0.33	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-0.36	0.33	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	0.21	0.33	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	0.64	0.34	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	-0.51	0.33	item	35	SF	secondary	TM	LO	no	Exp
Experiment1	0.13	0.33	item	35	SF	secondary	TM	LO	no	Exp
Corral et al. (2021)										
Experiment1	-0.11	0.19	item	11	SF	tertiary	STEM	HO	no	Exp
Experiment1	0.18	0.19	item	11	SF	tertiary	STEM	HO	no	Exp
Experiment2	0.33	0.19	item	11	EF	tertiary	STEM	HO	no	Exp
Experiment2	0.20	0.19	item	11	EF	tertiary	STEM	HO	no	Exp
Experiment3	0.59	0.25	sec.	172800	EF	tertiary	STEM	HO	no	Exp
Experiment3	0.43	0.24	sec.	172800	EF	tertiary	STEM	HO	no	Exp
Ding et al. (2024)										
Experiment1	0.13	0.31	sec.	23	SF	tertiary	GCS	HO	yes	Exp
Experiment1	-0.91	0.32	sec.	23	SF	tertiary	GCS	HO	yes	Exp
Experiment1	03	0.31	sec.	23	SF	tertiary	GCS	HO	yes	Exp
Experiment1	0.16	0.31	sec.	30	SF	tertiary	GCS	HO	yes	Exp
Experiment1	-18	0.33	sec.	30	SF	tertiary	GCS	HO	yes	Exp
Experiment1	07	0.31	sec.	30	SF	tertiary	GCS	HO	yes	Exp
Experiment1	-07	0.31	sec.	35	SF	tertiary	GCS	HO	yes	Exp

Table 6 continued from previous page

Study ID / Experiment	g	SE	Delay Unit	Delay Difference	Feedback Type	Educational Level	Learning Domain	Task Complexity	Prior Know. Adj.	Task Context
Experiment1	-06	0.31	sec.	35	SF	tertiary	GCS	HO	yes	Exp
Experiment1	0.13	0.31	sec.	35	SF	tertiary	GCS	HO	yes	Exp
Dunbar et al. (2017)										
Experiment1	-08	0.10	sec.	–	EF	tertiary	GCS	HO	no	Exp
Erdman et al. (2013)										
Experiment1	06	0.18	item	47	SF	tertiary	GCS	LO	no	Exp
Erdmann et al. (2022)										
Experiment1	0.67	0.22	item	–	EF	tertiary	LL	HO	no	Exp
Experiment1	0.24	0.21	item	–	EF	tertiary	LL	HO	no	Exp
Experiment1	-0.48	0.22	item	–	EF	tertiary	LL	HO	no	Exp
Experiment1	-05	0.21	item	–	EF	tertiary	LL	HO	no	Exp
Fyfe and Rittle-Johnson (2016)										
Experiment1	0.32	0.28	item	11	SF	primary	STEM	HO	no	CB
Fyfe (2016)										
Experiment2	0.24	0.24	item	11	SF	primary	STEM	HO	no	CB
Experiment2	04	0.24	item	11	SF	primary	STEM	HO	no	CB
Goda (2004)										
Experiment1	-05	0.21	item	19	EF	tertiary	LL	HO	no	CB
Experiment1	-02	0.21	item	19	EF	tertiary	LL	HO	no	CB
Guzmán-Muñoz et al. (2008)										
Experiment1	-0.64	0.30	item	27	SF	tertiary	GCS	LO	no	Exp
Experiment1	-0.44	0.29	item	27	SF	tertiary	GCS	LO	no	Exp
Hays et al. (2013)										
Experiment1	0.60	0.13	sec.	570	SF	tertiary	GCS	LO	no	Exp
Experiment2	0.37	0.15	sec.	570	SF	tertiary	GCS	LO	no	Exp
Experiment3	0.36	0.14	sec.	570	SF	tertiary	GCS	LO	no	Exp
Henderson (2021)										
Experiment1	0.16	0.43	item	9	SF	tertiary	LL	LO	no	CB
Henshaw (2011)										
Experiment1	-02	0.27	mixed	–	EF	tertiary	LL	HO	no	CB
Experiment1	-0.27	0.27	mixed	–	EF	tertiary	LL	HO	no	CB
Experiment1	0.31	0.27	item	39	EF	tertiary	LL	HO	no	CB

Table 6 continued from previous page

Study ID / Experiment	g	SE	Delay Unit	Delay Difference	Feedback Type	Educational Level	Learning Domain	Task Complexity	Prior Know. Adj.	Task Context
Experiment1	-0.24	0.27	item	39	EF	tertiary	LL	HO	no	CB
Iwaki et al. (2017)										
Experiment1	-0.24	0.17	item	59	SF	tertiary	LL	LO	no	Exp
Experiment1	0.11	0.17	item	59	SF	tertiary	LL	LO	no	Exp
Johnson et al. (2015)										
Experiment1	0.81	0.24	item	2	EF	secondary	STEM	HO	no	CB
Kourtali et al. (2023)										
Experiment1	-0.70	0.27	item	–	SF	secondary	LL	LO	yes	Exp
Experiment1	-0.17	0.26	item	–	SF	secondary	LL	LO	yes	Exp
E. H. Lavolette (2014)										
Experiment1	0.21	0.19	item	31	EF	tertiary	LL	LO	no	CB
Experiment1	-04	0.19	item	31	EF	tertiary	LL	LO	no	CB
Experiment1	0.33	0.19	item	31	SF	tertiary	LL	LO	no	CB
Experiment1	0.30	0.19	item	31	SF	tertiary	LL	LO	no	CB
E. Lavolette et al. (2015)										
Experiment1	06	0.34	sec.	–	EF	tertiary	LL	HO	no	CB
Lu, Sales, et al. (2021)										
Experiment1	0.29	08	sec.	604800	SF	tertiary	LL	LO	yes	CB
Experiment1	0.11	08	sec.	604800	SF	tertiary	LL	LO	yes	CB
Lu, Wei Wang, et al. (2023)										
Experiment1	0.74	0.30	item	5	EF	tertiary	LL	LO	no	CB
Metcalfe, Kornell, et al. (2009)										
Experiment1	-0.54	0.24	sec.	86400	SF	primary	LL	LO	no	CB
Experiment1	-1.29	0.31	sec.	86400	SF	primary	LL	LO	no	CB
Experiment2	00	0.21	sec.	86400	SF	tertiary	LL	LO	no	CB
Experiment2	-0.72	0.24	sec.	86400	SF	tertiary	LL	LO	no	CB
G. R. Morrison et al. (1995)										
Experiment1	-0.15	0.31	item	23	SF	tertiary	SS	HO	no	CB
Experiment1	-0.17	0.26	item	23	SF	tertiary	SS	HO	no	CB
Mullaney et al. (2014)										
Experiment1	0.20	0.19	sec.	4	SF	tertiary	GCS	LO	no	Exp
Experiment1	-0.54	0.20	sec.	4	SF	tertiary	GCS	LO	no	Exp

Table 6 continued from previous page

Study ID / Experiment	g	SE	Delay Unit	Delay Difference	Feedback Type	Educational Level	Learning Domain	Task Complexity	Prior Know. Adj.	Task Context
Experiment2	-02	0.17	sec.	4	SF	tertiary	GCS	LO	no	Exp
Experiment2	-0.12	0.17	sec.	4	SF	tertiary	GCS	LO	no	Exp
Experiment2	05	0.17	sec.	–	SF	tertiary	GCS	LO	no	Exp
Experiment2	-0.38	0.17	sec.	–	SF	tertiary	GCS	LO	no	Exp
Mullet et al. (2014)										
Experiment1	-0.85	0.42	sec.	604800	EF	tertiary	STEM	HO	no	CB
Experiment2	-0.29	0.17	sec.	604800	EF	tertiary	STEM	HO	no	CB
Experiment2	-0.65	0.18	sec.	604800	EF	tertiary	STEM	HO	no	CB
Nakata (2015)										
Experiment1	-05	0.10	item	–	SF	tertiary	LL	LO	no	CB
Experiment1	-02	0.10	item	–	SF	tertiary	LL	LO	no	CB
Experiment1	03	0.10	item	–	SF	tertiary	LL	LO	no	CB
Experiment1	-0.10	0.10	item	–	SF	tertiary	LL	LO	no	CB
Experiment1	0.12	0.10	item	–	SF	tertiary	LL	LO	no	CB
Experiment1	-03	0.10	item	–	SF	tertiary	LL	LO	no	CB
Nunn et al. (2024)										
Experiment1	0.10	0.23	sec.	5.5	SF	adult ed.	GCS	LO	no	Exp
Experiment1	-0.49	0.25	sec.	5.5	SF	adult ed.	GCS	LO	no	Exp
Opitz et al. (2011)										
Experiment1	1.10	0.19	sec.	1	SF	tertiary	LL	HO	no	Exp
Ryan et al. (2024)										
Experiment1	05	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Experiment1	09	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Experiment1	0.26	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Experiment1	-0.30	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Experiment1	0.36	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Experiment1	05	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Experiment1	00	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Experiment1	-08	0.16	item	8	EF	tertiary	STEM	HO	no	CB
Shintani et al. (2016)										
Experiment1	0.57	0.29	sec.	1800	SF	tertiary	LL	HO	no	CB
Experiment1	0.65	0.29	sec.	1800	SF	tertiary	LL	HO	no	CB

Table 6 continued from previous page

Study ID / Experiment	g	SE	Delay Unit	Delay Difference	Feedback Type	Educational Level	Learning Domain	Task Complexity	Prior Know. Adj.	Task Context
Shirah et al. (2023)										
Experiment1	0.22	0.19	item	11	SF	tertiary	STEM	HO	no	Exp
N. Sinha et al. (2015)										
Experiment1	-0.40	06	item	8	SF	tertiary	STEM	HO	no	CB
Experiment1	0.10	06	item	8	SF	tertiary	STEM	HO	no	CB
Experiment2	-0.60	0.20	item	19	SF	tertiary	STEM	HO	no	Exp
Experiment2	-0.27	0.19	item	19	SF	tertiary	STEM	HO	no	Exp
Experiment3	-0.36	0.13	item	19	SF	tertiary	STEM	HO	no	Exp
Experiment3	05	0.13	item	19	SF	tertiary	STEM	HO	no	Exp
Sitzman et al. (2014)										
Experiment2	02	0.13	sec.	1500	SF	tertiary	LL	LO	no	Exp
Experiment3	-09	0.18	sec.	1500	SF	tertiary	LL	LO	no	Exp
Smith et al. (2010)										
Experiment1	-05	0.10	sec.	480	SF	tertiary	GCS	LO	no	Exp
Experiment1	-0.49	0.10	sec.	480	SF	tertiary	GCS	LO	no	Exp
Experiment1	00	0.10	sec.	480	SF	tertiary	GCS	LO	no	Exp
Experiment1	-09	0.10	sec.	480	SF	tertiary	GCS	LO	no	Exp
Smits et al. (2008)										
Experiment1	0.39	0.29	item	4	SF	secondary	STEM	HO	no	CB
Experiment1	-0.42	0.36	item	4	SF	secondary	STEM	HO	no	CB
Experiment1	-08	0.28	item	4	EF	secondary	STEM	HO	no	CB
Experiment1	0.25	0.38	item	4	EF	secondary	STEM	HO	no	CB
Strong et al. (2019)										
Experiment1	-0.36	0.26	item	13	SF	tertiary	LL	LO	no	CB
Experiment1	0.23	0.26	item	13	SF	tertiary	LL	LO	no	CB
S. Tanaka et al. (2019)										
Experiment1	05	0.18	sec.	336	SF	tertiary	GCS	LO	no	Exp
Experiment1	0.13	0.18	sec.	336	SF	tertiary	GCS	LO	no	Exp
Experiment2	0.18	0.19	sec.	270	SF	tertiary	GCS	LO	no	Exp
Experiment2	0.28	0.19	sec.	624	SF	tertiary	GCS	LO	no	Exp
Experiment3a	0.16	0.19	sec.	1224	SF	tertiary	GCS	LO	no	Exp
Experiment3a	0.16	0.19	sec.	1224	SF	tertiary	GCS	LO	no	Exp

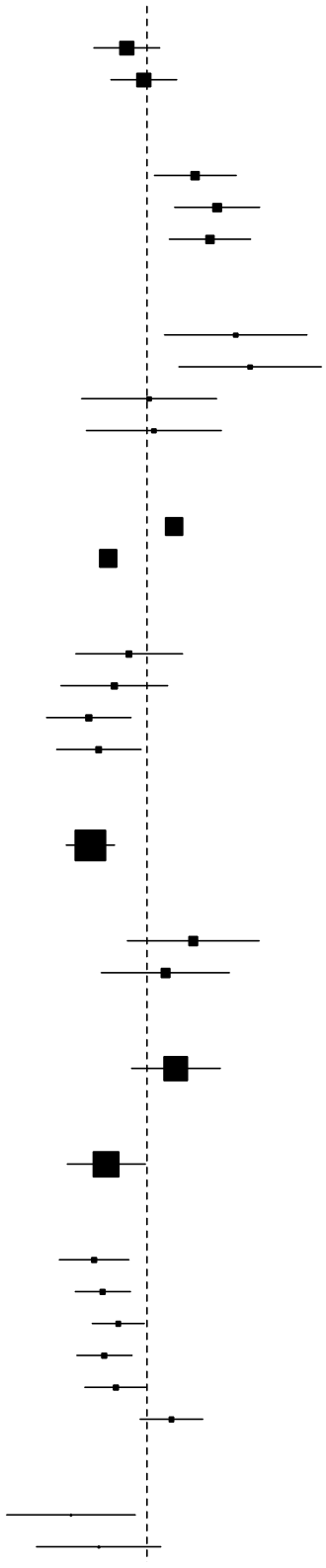
Table 6 continued from previous page

Study ID / Experiment	g	SE	Delay Unit	Delay Difference	Feedback Type	Educational Level	Learning Domain	Task Complexity	Prior Know. Adj.	Task Context
Experiment3b	-05	0.19	sec.	1200	SF	tertiary	GCS	LO	no	Exp
Experiment3b	0.68	0.21	sec.	1200	SF	tertiary	GCS	LO	no	Exp
Taxipulati et al. (2021)										
Experiment1	0.32	0.28	item	9	EF	tertiary	GCS	HO	no	Exp
Experiment1	0.51	0.28	item	9	SF	tertiary	GCS	HO	no	Exp
Van der Kleij, Eggen, et al. (2012)										
Experiment1	-0.14	0.20	item	29	EF	tertiary	SS	HO	no	Exp
Y. Wang et al. (2023)										
Experiment2	-17	0.38	sec.	86400	EF	secondary	RC	HO	no	Exp
Experiment2	-0.68	0.35	sec.	86400	EF	secondary	RC	HO	no	Exp
Yilmaz et al. (2019)										
Experiment1	1.27	0.41	item	11	SF	tertiary	LL	HO	yes	Exp
Experiment1	17	0.40	item	11	SF	tertiary	LL	HO	yes	Exp
Experiment1	00	0.37	item	11	SF	tertiary	LL	HO	yes	Exp
Experiment1	0.36	0.37	item	11	SF	tertiary	LL	HO	yes	Exp
Zawadzka et al. (2023)										
Experiment1	00	0.13	item	39	SF	tertiary	GCS	LO	no	Exp
Experiment1	00	0.13	item	39	SF	tertiary	GCS	LO	no	Exp
Experiment2	-0.11	0.13	item	39	SF	tertiary	GCS	LO	no	Exp

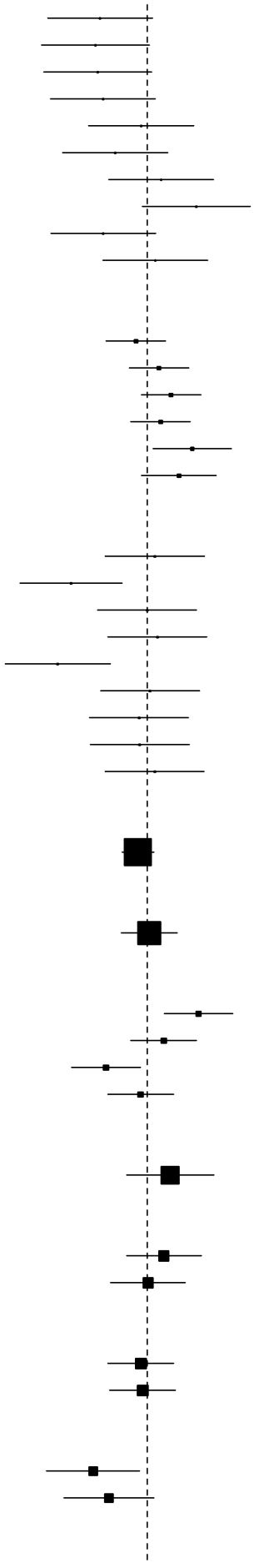
Notes. **Feedback Type:** SF = Simple Feedback, EF = Elaborated Feedback, TAF= Try-again Feedback. **Learning Domain:** GCS = General Cognitive Skills, LL = Language Learning, STEM = Science, Technology, Engineering, Math, SS = Social Sciences, TM = Text memory, RC= Reading Comprehension. **Task Complexity:** HO = Higher Order, LO = Lower Order. **Prior Knowledge Adjustment:** yes = Adjusted, no = Not Adjusted. **Learning Task Context:** Exp = Experimental, CB = Curriculum Based.

0.3 Forest Plot

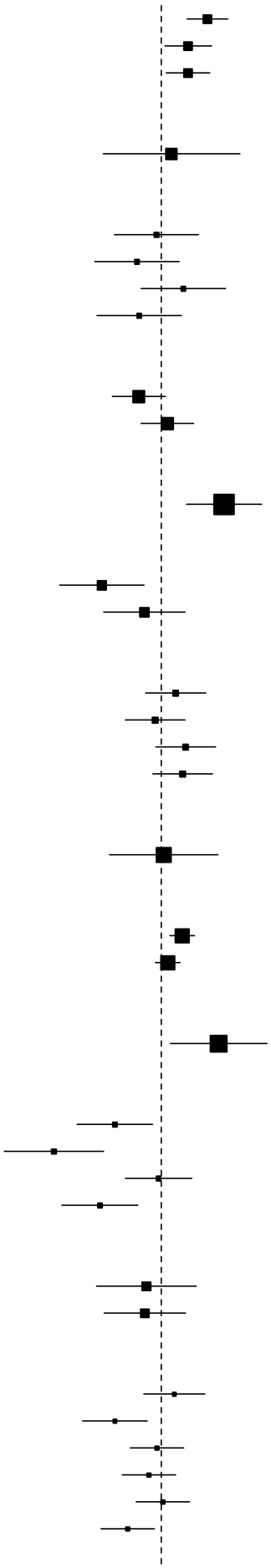
Forest Plot



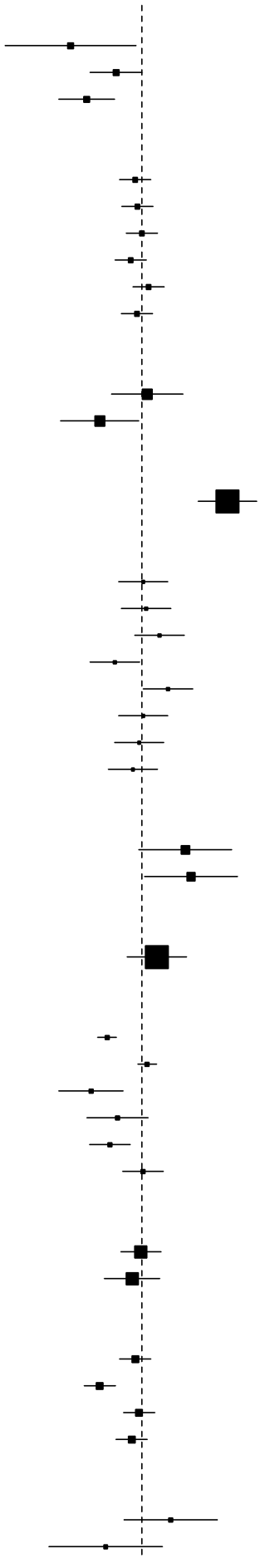
Studies	Weight	Effect Size
albrecht2023		
Experiment 1	3.318	-0.178
Experiment 1	3.318	0.000
aljabri2024		
Experiment 1	1.972	0.538
Experiment 1	1.972	0.768
Experiment 1	1.972	0.692
arroyo2018		
Experiment 1	0.973	0.962
Experiment 1	0.973	1.112
Experiment 1	0.973	0.055
Experiment 1	0.973	0.105
attali2017		
Experiment 1	4.111	0.316
Experiment 1	4.111	-0.372
butler2007		
Experiment 1	1.343	-0.154
Experiment 1	1.343	-0.308
Experiment 2	1.343	-0.575
Experiment 2	1.343	-0.471
butler2008		
Experiment 1	7.323	-0.558
canals2021		
Experiment 1	2.085	0.517
Experiment 1	2.085	0.227
candel2020		
Experiment 1	5.669	0.336
candel2021		
Experiment 1	6.121	-0.393
carpenter2011		
Experiment 1	1.146	-0.518
Experiment 2	1.146	-0.427
Experiment 2	1.146	-0.266
Experiment 2	1.146	-0.411
Experiment 3	1.146	-0.293
Experiment 3	1.146	0.290
clariana2000		
Experiment 1	0.358	-0.760
Experiment 1	0.358	-0.471



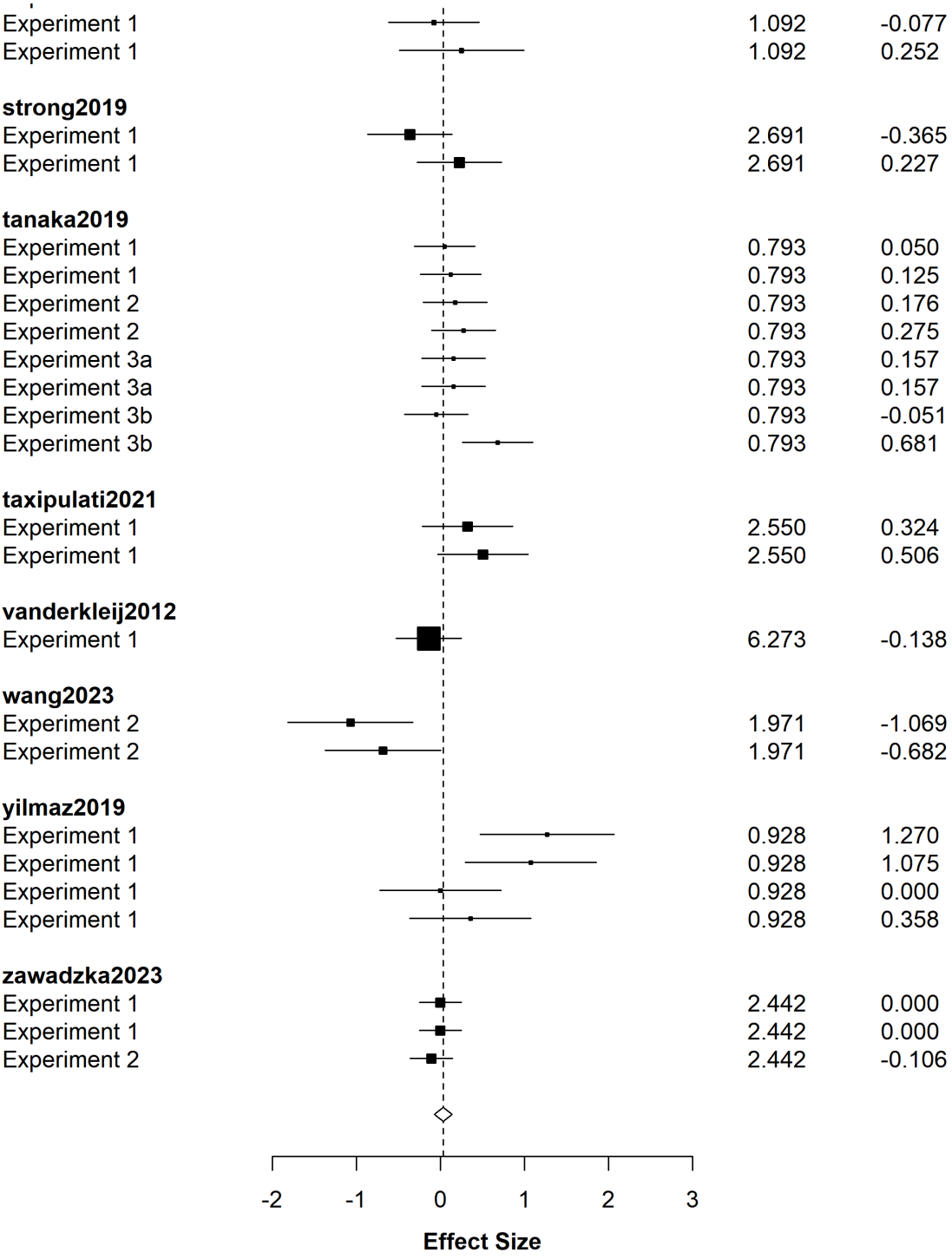
Experiment 1	0.358	-0.550
Experiment 1	0.358	-0.607
Experiment 1	0.358	-0.580
Experiment 1	0.358	-0.515
Experiment 1	0.358	-0.042
Experiment 1	0.358	-0.362
Experiment 1	0.358	0.206
Experiment 1	0.358	0.643
Experiment 1	0.358	-0.512
Experiment 1	0.358	0.133
corral2021		
Experiment 1	1.020	-0.107
Experiment 1	1.020	0.179
Experiment 2	1.020	0.330
Experiment 2	1.020	0.198
Experiment 3	1.020	0.592
Experiment 3	1.020	0.427
ding2024		
Experiment 1	0.504	0.128
Experiment 1	0.504	-0.910
Experiment 1	0.504	0.032
Experiment 1	0.504	0.159
Experiment 1	0.504	-1.075
Experiment 1	0.504	0.068
Experiment 1	0.504	-0.067
Experiment 1	0.504	-0.059
Experiment 1	0.504	0.127
dunbar2017		
Experiment 1	7.693	-0.080
erdman2013		
Experiment 1	6.595	0.058
erdman2022		
Experiment 1	1.511	0.670
Experiment 1	1.511	0.237
Experiment 1	1.511	-0.477
Experiment 1	1.511	-0.048
fyfe2016a		
Experiment 1	5.101	0.319
fyfe2016b		
Experiment 2	2.836	0.241
Experiment 2	2.836	0.043
goda2004		
Experiment 1	3.060	-0.045
Experiment 1	3.060	-0.023
guzmanmunoz2008		
Experiment 1	2.448	-0.637
Experiment 1	2.448	-0.442
hays2013		



Experiment 1	2.396	0.603
Experiment 2	2.396	0.366
Experiment 3	2.396	0.364
henderson2021		
Experiment 1	3.270	0.159
henshaw2011		
Experiment 1	1.310	-0.025
Experiment 1	1.310	-0.267
Experiment 1	1.310	0.307
Experiment 1	1.310	-0.237
iwaki2017		
Experiment 1	3.386	-0.244
Experiment 1	3.386	0.109
johnson2015		
Experiment 1	5.667	0.809
kourtali2023		
Experiment 1	2.654	-0.699
Experiment 1	2.654	-0.173
lavolette2014		
Experiment 1	1.606	0.211
Experiment 1	1.606	-0.040
Experiment 1	1.606	0.335
Experiment 1	1.606	0.298
lavolette2015		
Experiment 1	4.230	0.064
lu2021		
Experiment 1	3.963	0.293
Experiment 1	3.963	0.114
lu2023		
Experiment 1	4.698	0.742
metcalfe2009		
Experiment 1	1.356	-0.539
Experiment 1	1.356	-1.291
Experiment 2	1.356	0.000
Experiment 2	1.356	-0.723
morrison1995		
Experiment 1	2.473	-0.149
Experiment 1	2.473	-0.169
mullaney2014		
Experiment 1	1.095	0.196
Experiment 1	1.095	-0.538
Experiment 2	1.095	-0.019
Experiment 2	1.095	-0.119
Experiment 2	1.095	0.054
Experiment 2	1.095	-0.381



mullet2014		
Experiment 1	1.692	-0.851
Experiment 2	1.692	-0.285
Experiment 2	1.692	-0.652
nakata2015		
Experiment 1	1.283	-0.048
Experiment 1	1.283	-0.024
Experiment 1	1.283	0.035
Experiment 1	1.283	-0.104
Experiment 1	1.283	0.116
Experiment 1	1.283	-0.029
nunn2024		
Experiment 1	2.834	0.100
Experiment 1	2.834	-0.491
opitz2011		
Experiment 1	6.457	1.098
ryan2024		
Experiment 1	0.864	0.051
Experiment 1	0.864	0.087
Experiment 1	0.864	0.255
Experiment 1	0.864	-0.303
Experiment 1	0.864	0.357
Experiment 1	0.864	0.051
Experiment 1	0.864	0.000
Experiment 1	0.864	-0.077
shintani2016		
Experiment 1	2.416	0.573
Experiment 1	2.416	0.645
shirah2023		
Experiment 1	6.425	0.220
sinha2015		
Experiment 1	1.191	-0.399
Experiment 1	1.191	0.099
Experiment 2	1.191	-0.599
Experiment 2	1.191	-0.271
Experiment 3	1.191	-0.365
Experiment 3	1.191	0.049
sitzman2014		
Experiment 2	3.480	0.022
Experiment 3	3.480	-0.086
smith2010		
Experiment 1	1.925	-0.049
Experiment 1	1.925	-0.489
Experiment 1	1.925	0.000
Experiment 1	1.925	-0.093
smits2008		
Experiment 1	1.092	0.392
Experiment 1	1.092	-0.415

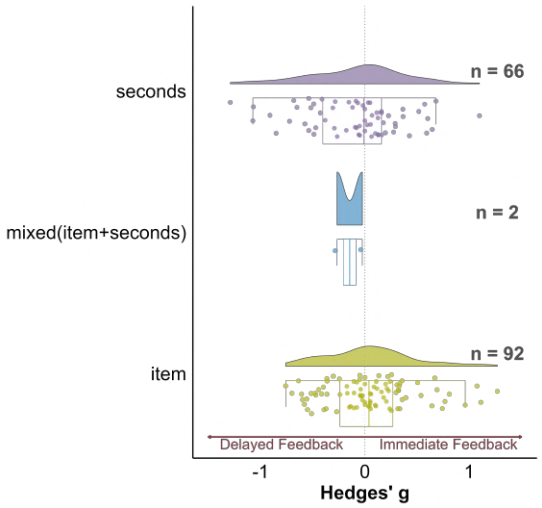

Experiment 1 1.092 -0.077
Experiment 1 1.092 0.252
strong2019
Experiment 1 2.691 -0.365
Experiment 1 2.691 0.227
tanaka2019
Experiment 1 0.793 0.050
Experiment 1 0.793 0.125
Experiment 2 0.793 0.176
Experiment 2 0.793 0.275
Experiment 3a 0.793 0.157
Experiment 3a 0.793 0.157
Experiment 3b 0.793 -0.051
Experiment 3b 0.793 0.681
taxipulati2021
Experiment 1 2.550 0.324
Experiment 1 2.550 0.506
vanderkleij2012
Experiment 1 6.273 -0.138
wang2023
Experiment 2 1.971 -1.069
Experiment 2 1.971 -0.682
yilmaz2019
Experiment 1 0.928 1.270
Experiment 1 0.928 1.075
Experiment 1 0.928 0.000
Experiment 1 0.928 0.358
zawadzka2023
Experiment 1 2.442 0.000
Experiment 1 2.442 0.000
Experiment 2 2.442 -0.106
-2 -1 0 1 2 3
Effect Size

0.4 Sensitivity Analysis

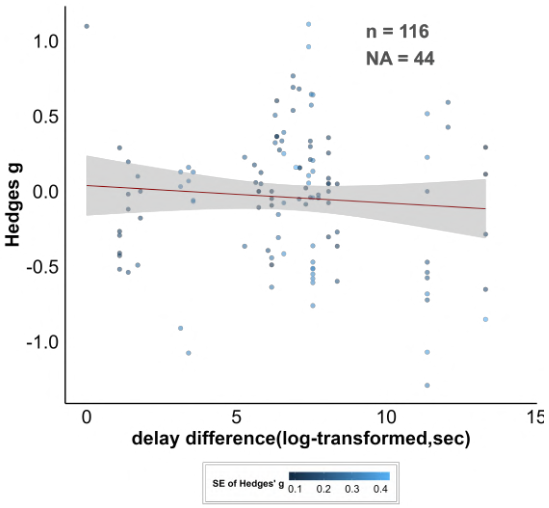
Table 7: Sensitivity Analysis for the RVE Model. Mean effect size (Hedges' g), standard error and estimated between-study variance (τ^2) depending on the assumed within-study effect size correlation (Rho) varying from 0 to 1.

	Rho = 0	Rho = 0.2	Rho = 0.4	Rho = 0.6	Rho = 0.8	Rho = 1
Mean effect size (g)	0.0341	0.0341	0.0341	0.0341	0.0341	0.0341
Std. Error	0.0524	0.0524	0.0524	0.0524	0.0524	0.0524
τ^2	0.1198	0.1199	0.1199	0.1200	0.1200	0.1201

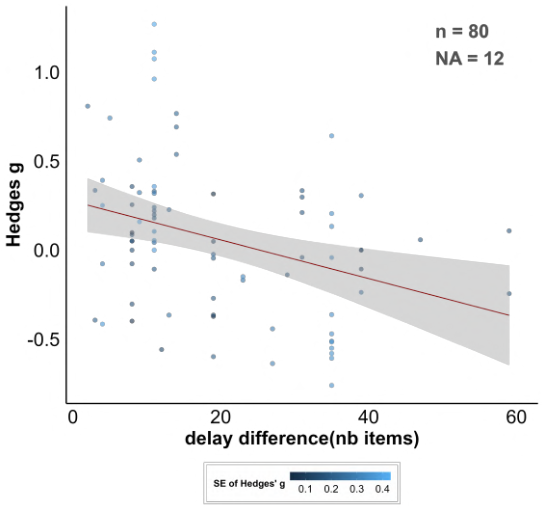
0.5 Distributions of Pre-registered Moderators



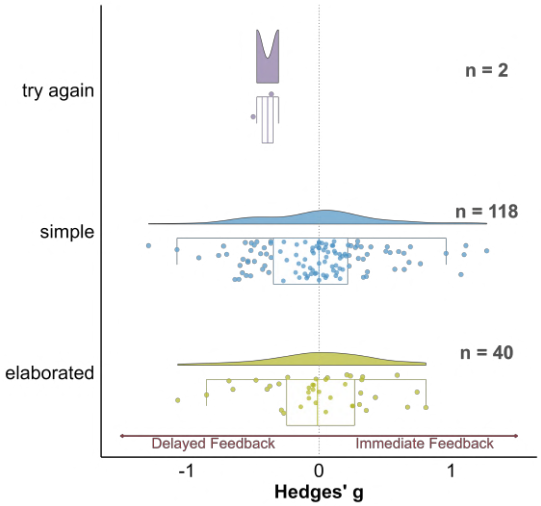
(a) Delay Unit



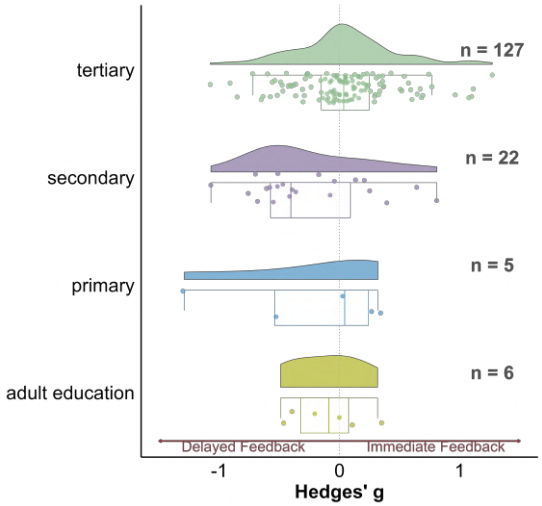
(b) Delay Difference-All Converted to Seconds



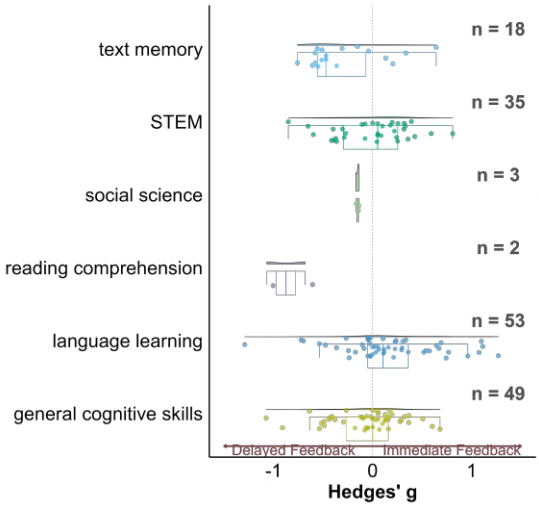
(c) Delay Difference (Items)



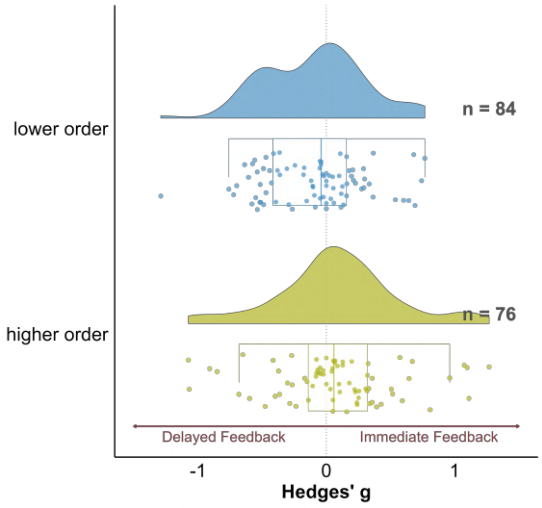
(d) Feedback Type



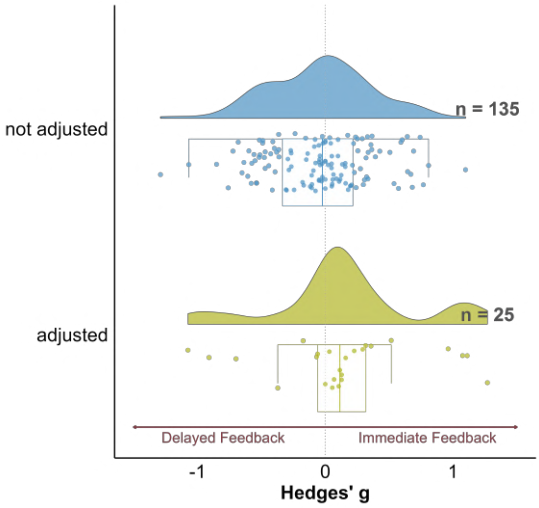
(e) Educational Level



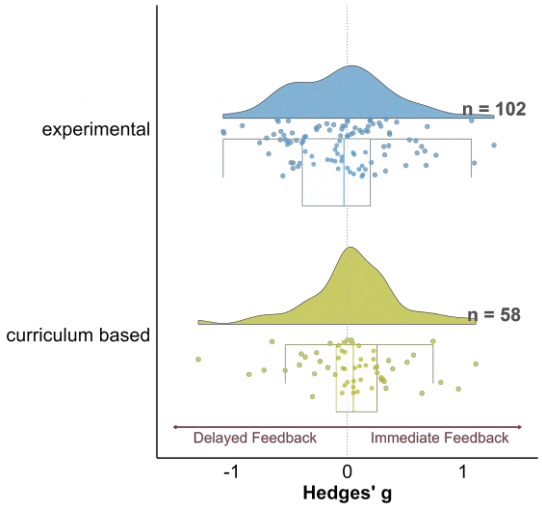
(f) Learning Domain



(g) Training Task Complexity



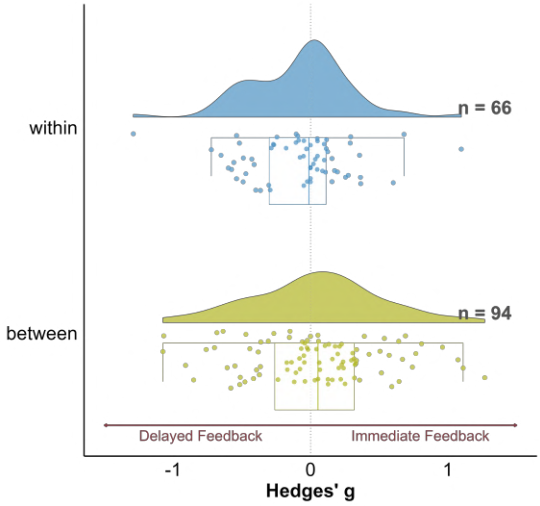
(h) Prior Knowledge Adjustment



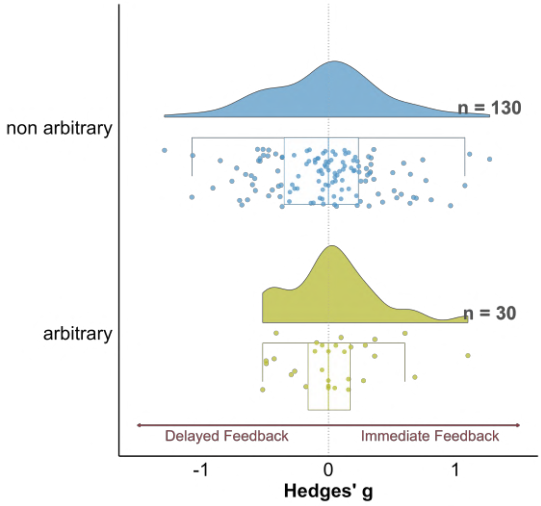
(i) Learning Task Context

Figure 4: Distribution of Pre-registered Moderators Across Effect Sizes

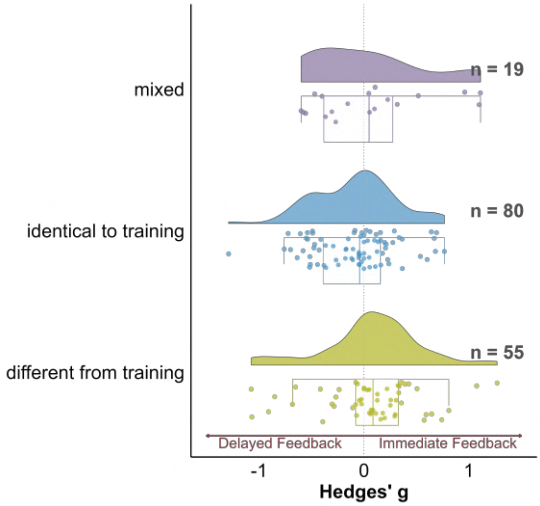
0.6 Distributions of Exploratory Moderators



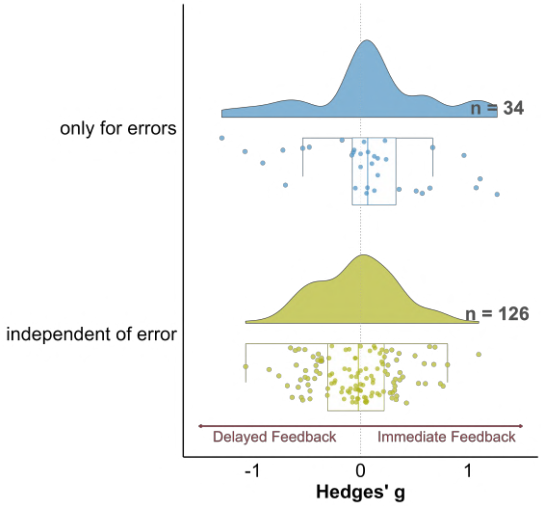
(a) Participant Design



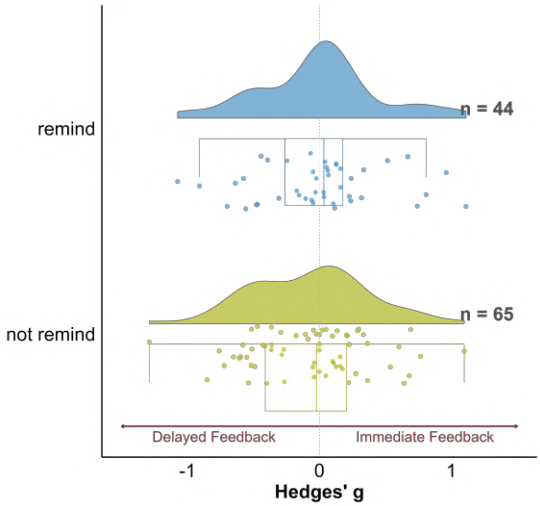
(b) Learning Task Type



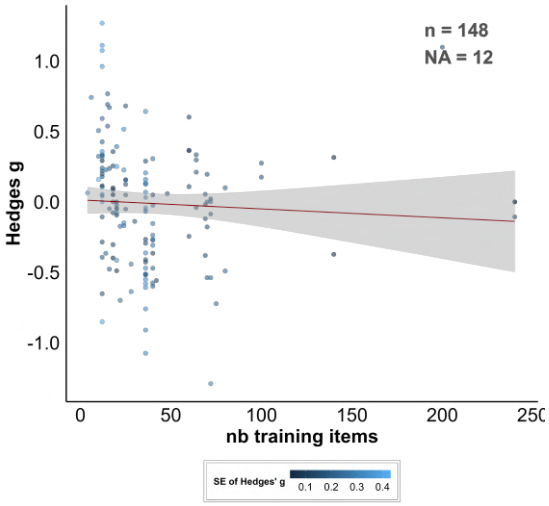
(c) Post-test Item Similarity



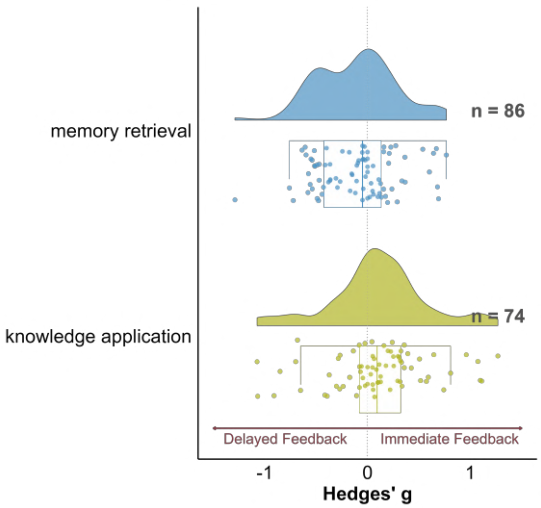
(d) Feedback Dependency on Errors



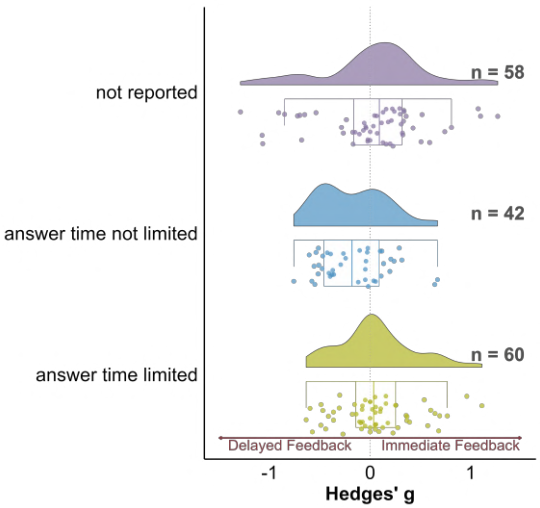
(e) Feedback answer reminder



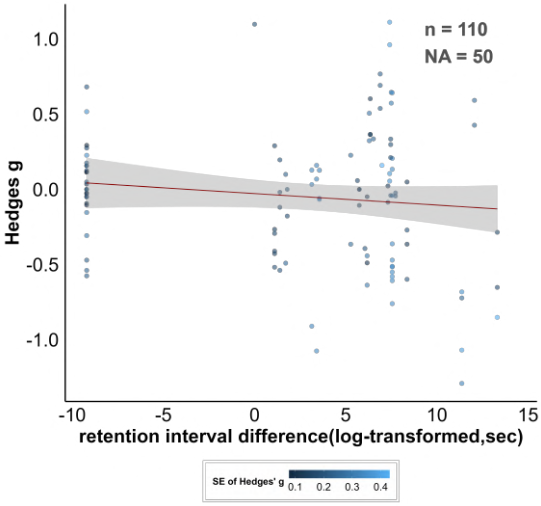
(f) Number of Training Items



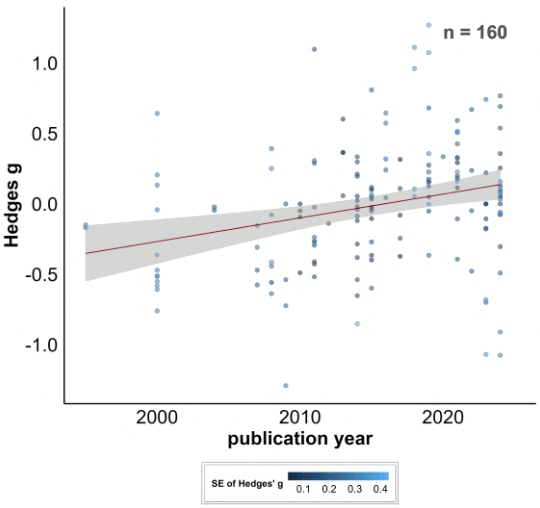
(g) Post-test Task



(h) Time Limitation for Responses



(i) Retention Interval Difference between Last Feedback and Post-test



(j) Publication Year

Figure 5: Distribution of Exploratory Moderators Across Effect Sizes

Chapter 5

The effect of feedback delay and initial answer recall on learning

This section constitutes the following manuscript:

Kandemir, E. N., Bakhtyar, M., Jouanot, F., Sanchez-Ayte, A., Palombi, O., & Ramus, F.
The Effect of Feedback Delay and Initial Answer Recall on Learning (in prep)
The preregistration for this experiment is available at this OSF link.

Contents

ABSTRACT Feedback is a critical instructional technique in educational settings, yet its optimal implementation remains a topic of debate, particularly concerning its timing and whether it reminds learners of their initial answers. While both immediate and delayed feedback have been shown to enhance learning under different conditions, recent theoretical and empirical work suggests that their effectiveness may depend on the presence or absence of initial answer recall during feedback. This study aimed to examine how feedback timing and initial answer recall interact to influence learning outcomes in a real-world, higher-order learning context: medical education. The experiment was conducted within a nationwide digital learning platform widely used by French medical students. Using a partially between-subjects design, each student trained on five medical specialties, each assigned to a distinct feedback condition crossing timing (immediate vs. delayed) and initial answer recall (recall vs. no recall), plus a control condition reflecting the platform's default feedback format. Learning outcomes were evaluated using post-test performance scores, statistically adjusted for pre-test performance. Analyses employed linear mixed-effects models controlling for student year, training intensity, and baseline ability. Due to the limited sample size, the results were inconclusive, with no statistically significant effects observed. Data collection is ongoing, and future analyses may yield more definitive insights to inform theory and optimize feedback delivery in digital medical education.

1 Introduction

Feedback is broadly defined as information provided by various sources on a learner's performance or understanding. This information enables learners to confirm, expand, modify, refine, or reorganize stored knowledge, which may include domain-specific knowledge, metacognitive skills, beliefs about oneself and tasks, or cognitive tactics and strategies (D. L. Butler et al., 1995; Mory, 2013).

Building upon this definition, extensive research across diverse learning environments and various academic disciplines has consistently demonstrated the significant positive impact of feedback on educational outcomes (A. C. Butler, Karpicke, et al., 2007; A. C. Butler and Roediger, 2008; Metcalfe, Kornell, et al., 2009; Mullet et al., 2014; Pashler, Nicholas J Cepeda, et al., 2005).

Despite the recognized benefits of feedback, its effectiveness can vary significantly. This variability highlights that its efficacy is not merely straightforward but contingent on multiple interrelated factors. These factors range from the intrinsic properties of the feedback, such as its type (Van der Kleij, Feskens, et al., 2015; M. Xu et al., 2023) and timing (Kulik et al., 1988; Mullet et al., 2014) to the characteristics of the learner, including their ability and motivation (Shute, 2008). Additional factors include the goals of the learning process (Van der Kleij, Feskens, et al., 2015; Golke et al., 2015), the method of feedback delivery (Swart et al., 2019; Hattie, 1999), the complexity of the task (Kluger et al., 1996), and the specific learning domain (Van der Kleij, Feskens, et al., 2015). Among these variables, two particularly complex and interrelated factors are the timing of feedback delivery and whether the feedback recalls the initial response of the learners. While the optimal timing for feedback remains an open question, recent theoretical and empirical research suggests that its effects may interact with initial answer recall. This article aims to address these intertwined questions by systematically investigating the combined effects of feedback timing and initial answer recall on learning outcomes.

1.1 Background literature

1.1.1 Feedback-Timing in the Computer-Assisted Learning Environments

Before delving into the ongoing debate about the optimal timing for providing feedback, it is essential to understand how feedback timing is categorized and how this categorization has evolved in technology-rich learning environments.

As described by Shute (2008), feedback timing within educational settings generally falls into two categories: immediate and delayed. Immediate feedback is provided directly after a learner completes a task or a problem-solving step or provides an answer. On the other hand, delayed feedback spans a broader timeline, ranging from a few hours to several days or even a week after task completion. This variability in the timeline for delayed feedback, combined with the practical challenges of providing truly immediate feedback in paper-based tasks, introduces a level of ambiguity. This ambiguity can make it challenging to consistently distinguish between immediate and delayed feedback across different studies. As a result, feedback that is considered delayed in one study may be classified as immediate in another. Mory (2013) suggested that

this ambiguity might contribute to the inconsistent findings regarding the impacts of immediate versus delayed feedback seen in educational research.

However, advances in educational technology have introduced more systematic approaches to categorize feedback timing within computer-assisted learning environments. These environments leverage technology to provide feedback that is not only truly immediate but also consistent and scalable, which offers clear advantages over traditional teaching methods (Arroyo et al., 2018; Ziegler, 2016). In these settings, delayed feedback is commonly referred to as all feedback that is not provided immediately after the learner gives an answer (Eggen et al., 2011). While not all computer-based studies adhere strictly to this definition, it provides a clearer distinction that enables more reliable comparisons across studies on immediate versus delayed feedback. Furthermore, educators and researchers emphasize the benefits of feedback in computer-assisted learning environments, emphasizing its potential to enhance learning experiences, provide strategic information effectively, and facilitate the construction of meaningful knowledge (Van Ginkel et al., 2019; Narciss, 2013; Fu et al., 2018).

1.1.2 Previous Feedback-Timing Studies

From a behaviorist perspective, feedback functions as a form of reinforcement (Skinner, 1958). Thus, for learning to occur effectively, feedback should be given immediately after the performance. A range of empirical studies supports this notion, demonstrating that immediate feedback tends to enhance learning and retention more effectively than delayed feedback (White, 1968; Arroyo et al., 2018; Li et al., 2016; Lu, Sales, et al., 2021; Mondigo et al., 2017).

However, several studies (A. C. Butler, Karpicke, et al., 2007; Carpenter and Vul, 2011; Kulhavy and R. C. Anderson, 1972; Smith et al., 2010; Guzmán-Muñoz et al., 2008) have shown that delayed feedback can improve learning more effectively, a phenomenon known as the delay-retention effect (Metcalfe, Kornell, et al., 2009; Mory, 2013).

In their seminal meta-analysis, Kulik et al. (1988) reviewed studies employing diverse methodologies and concluded that while delayed feedback can improve learning more effectively than immediate feedback, this advantage is generally limited to controlled, artificial laboratory settings with relatively simple stimuli. Additionally, Clariana et al. (2000) noted that the difficulty of the learning material itself might affect the optimal timing of feedback; easy items may benefit more from delayed feedback, whereas difficult items seem better learnt with immediate feedback, although these findings did not reach statistical significance. Further studies have shown that immediate feedback tends to support lower-order learning tasks like memorization, while delayed feedback is more effective for higher-order learning tasks that require knowledge application and transfer (Shute, 2008; Van der Kleij, Feskens, et al., 2015). Additionally, the interaction between feedback timing and a learner's ability has been highlighted (Nathan et al., 2002).

Recent meta-analytic findings (Author, 2025) highlight that the effects of feedback timing are not uniform but are moderated by several study and task characteristics. Based on a systematic analysis of 51 studies published between 1988 and 2024, the results showed that educational level, learning domain, post-test task type, and response time constraints significantly influence whether immediate or delayed feedback is more effective. Delayed feedback appeared partic-

ularly beneficial for secondary education students and tasks involving reading comprehension and memory recall, especially when learners were given unlimited time to respond. In contrast, immediate feedback tended to be more effective in tasks emphasizing knowledge application, particularly in tertiary education contexts. These findings suggest that the effectiveness of feedback timing is highly context-dependent and shaped by characteristics of both the learner and the task.

1.1.3 The Role of Initial Answer Recall in Learning

While educational research has extensively examined feedback timing, the cognitive processes that occur during feedback reception—particularly initial answer recall—deserve greater attention. In the context of the present study, initial answer recall refers to whether the student's original response is shown during the feedback phase. Research on memory and learning has shown that recalling prior errors can have powerful effects on subsequent learning. The hypercorrection effect, for example, suggests that recalling one's errors—especially when surprised by corrective information can lead to stronger memory encoding and better retention (A. C. Butler, Fazio, et al., 2011). Other studies have emphasized that generating and recalling errors can support deeper learning by engaging metacognitive monitoring and facilitating corrective processing (Grimaldi et al., 2012; Kornell, Hays, et al., 2009; Metcalfe and J. Xu, 2018). Conversely, the interference-perseveration hypothesis (Kulhavy and R. C. Anderson, 1972) argues that recalling incorrect answers during feedback may sometimes hinder learning by reinforcing erroneous information.

Despite its theoretical importance, initial answer recall remains inadequately investigated in feedback research. Recent meta-analytic work (Author, 2025) examined whether answer reminders during feedback moderated feedback timing effects. While no significant moderating effect emerged, this finding warrants cautious interpretation, as most included studies did not experimentally manipulate answer recall and only a few explicitly reported whether or how initial responses were made available during feedback. Consequently, targeted experimental research that explicitly manipulates initial answer recall is essential to understand how it interacts with feedback timing and to clarify their combined effects on learning outcomes.

1.1.4 Theoretical background

Given the diverse empirical results on feedback timing effects, it is expected that theoretical explanations would also vary. The main theoretical disagreements focus on two key issues: the defined function of feedback and the perceived role of errors in the learning process.

The theoretical explanations supporting the superiority of immediate feedback are rooted in learning theory, which views feedback as a reinforcement of correct responses. These models are based on the principles of operant conditioning, which suggest that feedback is most effective when given shortly after the desired response; the further the delay, the less effective the reinforcement becomes (Hull, 1952; Saltzman, 1951; Skinner, 1965).

Conversely, the delay-retention effect, which supports the use of delayed feedback, has been explained by two dominant theoretical accounts: the memory-based accounts (including the

dual-trace hypothesis and the distributed practice effect) and the interference-perseveration theory.

In contrast to conditioning-based models, memory-based accounts suggest that feedback acts not merely as reinforcement, but as an opportunity for further encoding or strengthening of memory traces. The dual-trace hypothesis (Kulik et al., 1988; Clariana et al., 2000) posits that delayed feedback enhances learning by creating two distinct encoding opportunities: one during the initial response and another during feedback. These separate encoding events are thought to reinforce correct responses.

A related but distinct memory-based explanation is the distributed practice effect. According to this perspective, delayed feedback does not enhance learning because of the delay itself, but because it provides an additional spaced-retrieval opportunity. Research consistently shows that distributing retrieval attempts over time, rather than massing them together, substantially improves memory retention (Nicholas J Cepeda et al., 2009; Nicholas J. Cepeda et al., 2008). Because delayed feedback introduces a temporal gap from the original response, it naturally creates conditions for spaced retrieval, thereby enhancing memory independently of feedback timing per se (A. C. Butler, Karpicke, et al., 2007; Metcalfe, Kornell, et al., 2009). Some empirical support for this interpretation has been found (Smith et al., 2010). However, it is important to note that the distributed practice effect primarily benefits initially correct responses, as it strengthens existing memory traces rather than correcting errors (A. C. Butler, Karpicke, et al., 2007).

Conversely, the interference-perseveration theory (Kulhavy and R. C. Anderson, 1972) offers an explanation for the efficacy of delayed feedback specifically for the correction of initial errors. According to this theory, immediate feedback following an initially incorrect response can lead to an interference between their incorrect answer and the correct one. This confusion occurs because the erroneous response is assumed to still be active in their working memory, potentially resulting in learning the incorrect information. In contrast, providing feedback after a delay allows time for the initial error to be forgotten, reducing the likelihood of interference, and thus improving the learning process (Carpenter and Vul, 2011; Mory, 2013). Fundamentally, the theory is based on a few core assumptions: (a) learners make errors, (b) errors hinder learning, (c) feedback serves to correct these errors, and (d) delaying feedback improves error correction by allowing the error to be forgotten.

A substantial body of research supports this theory, showing that a higher proportion of initial errors are corrected on final tests when feedback is delayed rather than immediate (Kulhavy, 1977; Kulhavy and R. C. Anderson, 1972; Phye et al., 1989; Surber et al., 1975). These findings align with established memory theories, which made the interference-perseveration hypothesis a dominant explanation for the delay-retention effect (A. C. Butler and Roediger, 2008; Clariana et al., 2000; Kulik et al., 1988; Swindell et al., 1993).

However, there is also a growing body of research that challenges the assumptions of the interference-perseveration theory, presenting evidence that is inconsistent with its predictions. Firstly, the interference-perseveration theory assumes that learners make frequent errors, which limits its relevance in situations where few errors are made during learning. Metcalfe, Kornell,

et al. (2009) suggested that the effects of immediate and delayed feedback may depend on the frequency of errors.

Secondly, the theory assumes that errors hinder learning. However, studies have shown that generating errors can actually enhance the learning of corrective feedback (e.g., Grimaldi et al. (2012), Kornell, Hays, et al. (2009), and Metcalfe and J. Xu (2018)). A. C. Butler, Fazio, et al. (2011) also found that recalling previously generated errors can reinforce memory for corrective feedback rather than interfering with learning the correct answers, further challenging the theory's assumption that errors have a purely negative impact on learning.

From a more methodological perspective, the design of experiments supporting the interference perseveration theory often failed to control for the retention intervals between feedback conditions (e.g., A. C. Butler, Karpicke, et al. (2007), Kulhavy and R. C. Anderson (1972), O'neill et al. (1976), and Swindell et al. (1993)). This lack of control creates a potential confound, as delayed feedback is often given closer to the post-test than immediate feedback. Given that shorter intervals between retrieval and testing generally result in better memory performance than longer intervals (Nicholas J. Cepeda et al., 2008; Metcalfe, Kornell, et al., 2009), Metcalfe, Kornell, et al. (2009) suggest that the observed delay-retention effect may be at least partially attributable to the proximity of feedback to the test rather than the interference perseveration hypothesis alone. Studies by Metcalfe, Kornell, et al. (2009) and Nakata (2015) have shown that when the interval between feedback and testing is controlled, the timing of feedback may not significantly impact learning outcomes.

Moreover, the interference-perseveration theory assumes that errors remain active in working memory only with immediate feedback, leading to interference. Yet, it does not account for the possibility of errors being retrieved from long-term memory when feedback is delayed. If the interference-perseveration hypothesis was correct, then providing learners with reminders of their initial errors before feedback—whether immediate or delayed, should cause proactive interference, leading to worse performance. In other words, being reminded of past errors should disrupt the acquisition of correct feedback regardless of when the feedback is given. Iwaki et al. (2017) tested this assumption by using visual cues to remind participants of their errors, manipulating both feedback timing and the presence of error cues. Contrary to the theory's prediction, they found no significant difference between groups; in fact, participants who were reminded of their errors performed better. This further supports the beneficial impact of error recall and suggests that the interference-perseveration theory can only explain the delay retention effect in situations where feedback or cues do not remind initial errors.

Finally, the interference-perseveration theory tends to be more relevant for lower-order learning tasks, where errors in working memory can disrupt learning by interfering with the acquisition of correct responses (L. W. Anderson et al., 2001; Bloom et al., 1964). Conversely, for higher-order learning tasks that involve applying knowledge from long-term memory in new contexts, errors in working memory may not necessarily hinder learning. Thus, this theory might not fully account for learning in higher-order learning contexts, which require more complex feedback beyond providing simple corrective information (Eggen et al., 2011).

1.2 Limitations of previous studies

In addition to inconsistencies between previous studies regarding feedback timing and theoretical frameworks, there are also significant gaps in the literature on feedback timing.

First, most research on feedback timing in computer-based assessments has focused on lower-order outcomes, such as rote memorization (Eggen et al., 2011; Van der Kleij, Feskens, et al., 2015), with few studies examining its effects or testing the interference-perseveration theory in the context of higher-order learning tasks.

Additionally, in studies where immediate feedback was found to be more effective, delayed feedback was often provided through methods such as a written list of errors with corrections (Arroyo et al., 2018) or by repeating the learner's error before presenting the correction (Li et al., 2016). However, studies supporting delayed feedback often focused on lower-level tasks and did not prompt learners to recall their initial answers before or during feedback (Carpenter and Vul, 2011; A. C. Butler, Karpicke, et al., 2007). This suggests that initial answer recall and feedback timing have not been properly decoupled and may interact, yet there are very few studies exploring this interaction (e.g., Iwaki et al. (2017)).

Taken together, these gaps emphasize the need for further investigation. Specifically, research should explore how feedback timing interacts with initial answer recall in higher-order learning tasks to more effectively test the interference-perseveration theory and its underlying assumptions.

1.3 The present study

Building on this research gap, the present study aims to examine the effects of feedback timing, initial answer recall, and their interaction on medical students' learning outcomes in the context of long-term, real-life medical education training. Medical education is a higher-order learning domain requiring deep understanding, long-term retention, and the ability to apply knowledge in practical settings.

The following research question guided our research:

- **RQ:** How do feedback timing (immediate vs. delayed) and the recall of initial answers interact to influence learning outcomes in higher-level learning tasks, such as medical education?

Based on the existing literature, three primary hypotheses have been formulated:

- **H1:** The effectiveness of feedback for learning is influenced by its timing (immediate vs. delayed).
- **H2:** The effectiveness of feedback for learning is influenced by whether initial answers are recalled at the time of feedback.
- **H3:** The interaction between feedback timing and initial answer recall is non-additive.

Figure 5.1: Schematic illustration of the predicted interaction patterns between feedback timing and initial answer recall.

Based on the recent meta-analysis (Author, 2025) and the characteristics of the present study (tertiary education participants, STEM learning domain, knowledge application tasks, and no response time constraints), we anticipate that the effect of feedback timing on learning outcomes may be limited. While the absence of response time constraints could slightly favor delayed feedback, the nature of the tasks may conversely benefit from immediate feedback. Given these competing influences, we expect any main effect of feedback timing to be small and potentially nonsignificant in this context.

Beyond this main effect, two theoretical frameworks make opposing predictions about the interaction between feedback timing and initial answer recall. According to interference-perseveration theory, delayed feedback is predicted to yield better learning than immediate feedback when students' initial answers are not recalled. When initial answers are recalled, the theory predicts no difference between immediate and delayed feedback, due to similar interference in both conditions. Thus, this theory predicts a feedback timing x initial answer recall interaction, with the highest performance in the delayed feedback without recall condition.

Alternative accounts emphasizing the benefits of error recall predict better learning outcomes when initial answers are recalled. No significant difference is predicted between immediate and delayed feedback when answers are recalled. However, when answers are not recalled, immediate feedback is expected to lead to slightly better learning than delayed feedback, since errors may remain in working memory. This framework, therefore, also predicts a feedback timing x initial answer recall interaction, but with the lowest performance in the delayed feedback without initial answer recall condition. These predicted interaction patterns are schematically illustrated in Figure 5.1.

To test the study's hypotheses, a within-subject experimental design was implemented in a computer-assisted learning system used by medical students. The two independent variables were feedback timing (immediate vs. delayed) and initial answer recall (recall vs. no recall), assigned across the five selected medical specialties for each participant. Although the study design exposed all participants to each experimental condition, the final analysis was conducted on a partially between-subject dataset due to incomplete data in some conditions and the constraints of a limited sample size.

2 Methods

2.1 Participants

After applying the inclusion criteria, (detailed in Section 2.5), participants were 20 medical students from the 2nd to 6th years recruited from 13 different French universities. Recruitment took place via the Banque Nationale d'Entrainement (BNE) experimental module of the digital learning system provided by Université Numérique en Santé et en Sport (UNESS). The study was advertised through medical faculties and students' associations as an opportunity to test an experimental version of the BNE, and therefore contribute to education research and to the improvement of the platform while training for their exams. Participation was voluntary and unpaid, with students free to choose to use either the experimental module or to continue with the standard BNE platform throughout the experiment. The study was approved by the institutional review board of Université Paris Cité (IRB #00012019-51 and amendment #00012023-17). All participants received full information on the protocol and provided informed consent.

2.2 Materials

2.2.1 The Online Learning Environment

BNE, provided by UNESS, is a digital learning system extensively utilized by over $8,800$ medical students from all French universities each academic year. It is primarily designed for training through multiple-choice questions. The platform houses a comprehensive question bank with approximately $2,313,023$ multiple-choice questions covering 31 medical specialties, including past exam questions and those specifically created for training purposes.

The training tests on BNE are categorized into three main types:

1. **Isolated Questions Test (IQ):** Classic multiple-choice questions.
2. **Progressive File Test (PF):** This test consists of roughly 15 interconnected multiple-choice questions, each revealing new medical information about a patient and simulating a realistic medical case.
3. **Critical Article Reading Test (CAR):** Questions are based on an article, testing the ability to integrate and apply information from written sources.

The standard BNE platform includes a training generation rubric that lets learners customize their sessions by selecting the type of test, the number of questions, and the medical specialties they want to focus on. In each session, the actual questions are randomly chosen from a larger question bank. Additionally, the platform incorporates a default feedback system; following each test, students receive corrective feedback on their answers' accuracy. This feedback not only highlights whether responses were correct but also displays the students' initial choices and, for a subset of questions, provides detailed explanations regarding the correctness of each answer.

2.2.2 Experimental Question Bank

The experimental question bank, a selectively curated subset of the main question bank, adhered to stringent criteria to ensure the highest quality and relevance:

- Only questions rated above 4 (out of 5) by students were included.
- The specialty of toxicology was excluded due to an insufficient number of questions to support robust analysis.
- The sequence of questions within each medical specialty and question type was randomised and fixed once and for all, in order for all students to take the same questions in the same order.

Overall, the experimental module included a total of 35,348 questions, distributed across various medical specialties as shown in Figure 5.2.

2.3 Experimental Design

The experiment employed a within-subjects design across different medical specialties.

2.3.1 Conditions

We manipulated two parameters to define the experimental conditions: the timing of the feedback and whether the initial answer was recalled during the feedback.

Feedback Timing:

- **Immediate Feedback:** Corrective feedback was provided immediately after each question.
- **Delayed Feedback:** Feedback was delivered at the end of each series of 15 questions. For DP/LCA, corrections occur at the end of each test, aligning with the default BNE mode where DP and LCA tests typically comprise approximately 15 questions. For the Isolated Questions (IQ) test—where the number of questions is normally flexible—test length was fixed at 15 questions in the delayed feedback conditions to ensure consistent feedback timing across test formats.

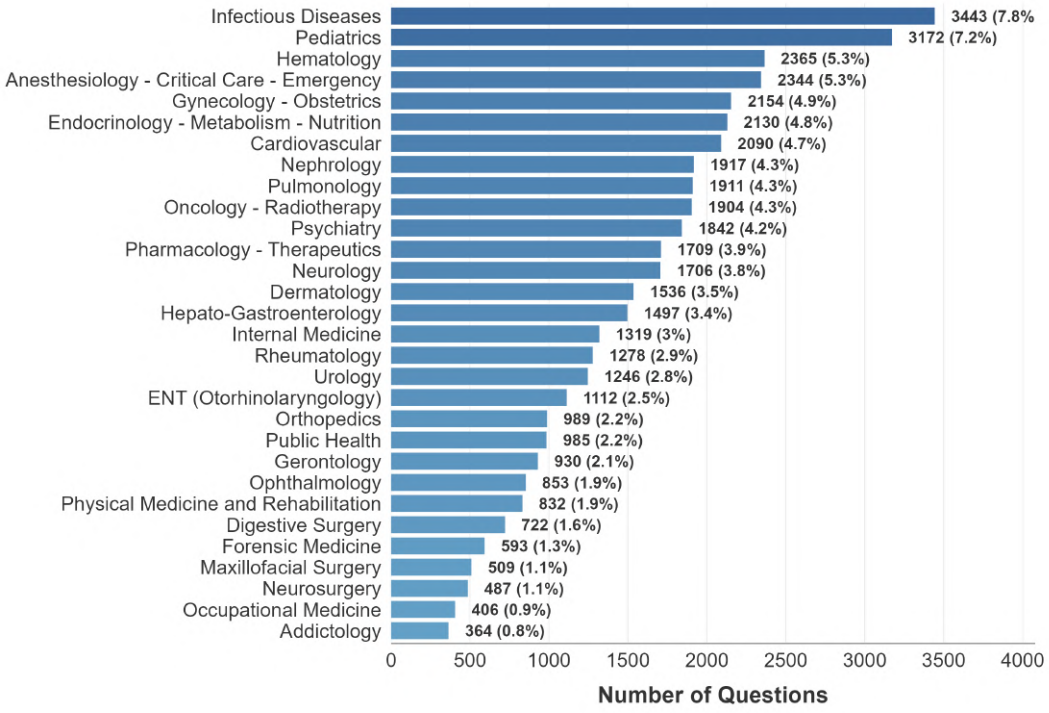


Figure 5.2: Distribution of the total number of questions by medical specialty, across all test types (QI, DP, and LCA)

Initial Answer Recall:

- **Initial Answer Recall (default mode of the BNE):** When students receive feedback, their initially selected options are displayed alongside the correct and incorrect options.
- **No Initial Answer Recall:** When students receive feedback, their initially selected options are hidden, showing only the correct and incorrect options.

Crossing these two parameters (feedback timing and initial answer recall) resulted in four experimental conditions (Table 5.1). An additional control condition was added, following the default BNE feedback format, where feedback is provided at the end of the test. Feedback delay is therefore not strictly controlled, as it varies depending on the question type and on the number of questions included in the test. The participant's initial answer is systematically recalled during feedback. This condition was included as a baseline to compare the experimental feedback manipulations against the standard learning experience on the platform.

For each participant, the five experimental conditions were pseudo-randomly assigned to the five chosen specialties. This assignment was managed by an algorithm designed to ensure that, across all participants, each specialty was assigned the 5 conditions a similar number of times, thus ensuring proper counterbalancing of conditions across specialties.

Condition	Feedback Timing	Initial Answer Recall
1. Control	Variable	Recall
2. Experimental	Delayed (15 questions)	No Recall
3. Experimental	Delayed (15 questions)	Recall
4. Experimental	Immediate	No Recall
5. Experimental	Immediate	Recall

Table 5.1: Experimental feedback conditions, defined by feedback timing and whether the student's initial answer is recalled during feedback. In the control condition, feedback timing varies with test format and is not experimentally controlled.

2.3.2 Experimental Module

For the purpose of this experiment, an experimental version of the BNE was designed and implemented within the digital learning system. This module introduced several significant modifications from the standard BNE platform:

- **Specialty Selection:** Each participant was required to select only five medical specialties from a list of 30, that they wished to particularly train on over the 3 months of the experiment. This flexibility was meant to accommodate the fact that participants were studying in different years in different universities, and therefore were at different stages of the medical curriculum. The five chosen specialties were then the only ones accessible in their training generation rubric. Should they want to train on other specialties that were not initially chosen, they were free to do so on the standard (non-experimental) version of the BNE.
- **Feedback Conditions:** For each participant, each chosen specialty was allocated to one of the five conditions, and was therefore associated with the corresponding type of feedback, as described in the experimental design section.
- **Question Presentation:** Contrary to the standard module's random question presentation, the experimental module had all participants within the same specialty receiving questions in a fixed sequence from the experimental question bank.

2.4 Procedure

The experiment ran from February 6th to May 11th, 2025. Upon registration in the experiment, participants completed a questionnaire providing demographic information (gender, university, and year of study) and detailing their use of the BNE platform and other resources for studying

medical specialties. Participants then selected five medical specialties to focus on during the experiment. These specialties were assigned the experimental feedback conditions as indicated above, and the study was restricted to these five specialties for each participant.

The experiment consisted of three phases:

Pre-test: Participants first completed a mandatory pre-test in each of their five selected specialties. Each pre-test consisted of the first 15 isolated questions from each specialty's question list. To ensure a neutral baseline assessment, the default BNE feedback format was employed for all pre-tests, irrespective of the experimental condition assigned to each student-specialty pair. Participants moved on to the training phase only after completing all five pre-tests. A summary report of their pre-test performance was made available to students.

Training: After completing the pre-test phase, participants entered a training period lasting approximately 2.5 months, from February 6th to April 25th, 2025. In each training session, the student selected the desired specialty and number of questions of each type, and the system sequentially presented new questions from the pre-generated question list. This ensured that all students who chose the same specialty trained on identical questions in a fixed order, thereby reducing irrelevant variability. The more a student trained in a specialty, the further they progressed through its question list. Feedback was provided according to the experimental condition assigned to each unique student-specialty pair.

Post-test: The post-test phase opened on April 4th and remained available until May 11th, 2025. This created a three-week overlap during which students could continue training while also beginning the post-tests. This design allowed students who felt ready in certain specialties to proceed directly to the post-test, while still giving them the opportunity to continue training in other specialties. After April 25th, training access was closed to prompt all remaining participants to complete their post-tests. Each participant was required to complete a post-test for each of their five selected specialties, within this one-month window. The post-test question lists were predetermined for each specialty and consisted of 45 isolated questions: the 15 questions from the pre-test, 15 questions selected from those presented during the training phase, and 15 entirely new questions (the final 15 questions from the specialty's question list, reserved exclusively for the post-test). The selection and order of questions were standardized for all students within each specialty. After completing each post-test, participants received feedback using the default BNE format.

Throughout the experiment, participation was entirely self-initiated and self-paced. Each student chose when, how long, and in which specialty they would train. They were sent periodic reminders of the availability of the experimental module and of the closing dates of the training and post-test phases. We had no control and no knowledge of any training the students did outside the experimental module.

Following the post-tests, participants completed an end-of-experiment questionnaire regarding their feedback preferences (timing and recall of initial answers), the time they spent training

in each specialty (using the experimental module, the standard BNE, and external resources), and their perception of the feedback manipulations' impact. Upon completion of the questionnaire, participants received a summary report detailing their pre-test and post-test performance and improvement for each specialty, marking the conclusion of the experiment.

2.5 Inclusion Criteria and Deviations from Pre-registration

To ensure valid and interpretable results, the study pre-registration specified a series of inclusion criteria for each phase of the experiment. However, due to lower-than-expected student engagement and the limited timeframe of the experiment, some of these criteria were relaxed in the final analysis. Below, we summarize both the preregistered and applied criteria, and explain the rationale behind any deviations.

Pre-test Phase

Preregistered:

- Inclusion required completion of all five pre-tests—one for each of the participant's selected specialties.

Applied:

- This criterion was maintained. Only students who completed all five pre-tests were included in the analysis. Completion of the five pre-tests was a prerequisite to accessing the training phase within the experimental module.

Training Phase

Preregistered:

- Participants had to complete at least 60 training questions in at least two of their selected specialties. This threshold was chosen to ensure sufficient exposure to the assigned feedback conditions and to enable within-subject comparisons across conditions.

Applied:

- The threshold was reduced to 15 questions per student–specialty pair. This filter was applied only to specialties assigned to an experimental condition (i.e., Conditions 2–5). The relaxation was necessary due to lower-than-expected engagement levels and allowed inclusion of partially engaged but still informative student–specialty pairs, even when data from only one condition were available.

Post-test Phase

Preregistered:

- Inclusion required completion of post-tests in at least two specialties in which participants had completed a minimum of 60 training questions. This was intended to support within-subject comparisons across feedback conditions. The preregistration did not specify what constituted post-test completion.

Applied:

- A student–specialty pair was retained if at least one post-test consisting of 45 isolated questions (IQ format) was completed in a specialty assigned to an experimental condition (Conditions 2–5). The IQ post-test was always the first test available and had to be completed before access to subsequent formats (e.g., DP). This decision allowed us to include participants who had not taken the post-test progressive file test.

Final Cleanup

Preregistered:

- Only students in years 3 to 5 of medical school were eligible.
- All five conditions (including the control condition) were to be included in the analysis.

Applied:

- Students in year 2 were also included in the final analysis. This adjustment improved statistical power and reflected the fact that some year 2 students were already studying the selected specialties.
- The control condition (Condition 1) was excluded from the final model. This condition involved default platform feedback behavior (variable timing, mandatory recall) and was not experimentally manipulated. Including it would have introduced interpretational ambiguity. The analysis was therefore restricted to the four experimental conditions that fully crossed feedback timing and recall. All inclusion criteria (e.g., training and post-test thresholds) were applied exclusively to student–specialty pairs under experimental conditions.

2.6 Variables and Measures

Dependent Variable: The primary outcome variable was post-test performance in each specialty, measured using the official scoring system of the platform, which assigns a score to each multiple-choice question answer based on the number of classification errors made by the student, referred to as discordances. A discordance occurs when a student ticks a choice that is incorrect, or fails to tick a choice that is correct. The overall question score is determined as follows: 1 point if there are no discordances, 0.5 points for one discordance, 0.2 points for two discordances, and 0 points if there are three or more.

Beyond this general rule, the scoring system also incorporates a mechanism to account for particularly critical errors. Some questions contain answer options flagged as essential or unacceptable. These labels indicate that certain choices are more important than others in terms of medical knowledge. For instance, failing to identify an essential choice—one that reflects essential clinical understanding—suggests a major knowledge gap. Similarly, selecting an unacceptable choice—one that reveals a dangerous misconception—also signals a serious error. In either case, the student's response is scored as 0 for the entire question, regardless of the number of other correct or incorrect selections. This scoring formula reflects the pedagogical emphasis and grading logic used within the BNE platform and was applied consistently across both pre-test and post-test assessments.

Independent Variables: The two manipulated variables were feedback timing and initial answer recall, as defined in the experimental design section. They were coded based on the feedback condition assigned to each participant–specialty pair.

Covariates:

- Pre-test Performance in each specialty: scored using the same scoring system as the post-test.
- Number of Training Questions: represented the total number of questions completed by each participant in a given specialty during the training phase.
- Student Year: recorded as an ordinal variable corresponding to the participant's current academic level (2nd, 3rd, 4th, 5th, or 6th year).

2.7 Analysis

The analyses employed linear mixed-effects models (LMMs) to accommodate the hierarchical structure of the data using the `lme4` package (Version 1.1-31) (Bates et al., 2015) in R (Version 4.2.2), applying the restricted maximum likelihood (REML) estimation method.

The primary analysis examined the main effects and interaction between feedback timing (immediate vs. delayed) and initial answer recall (recall vs. no recall) on student learning outcomes, operationalized as post-test performance. Covariates included pre-test performance, student year, and the total number of training questions completed within the specialty, as each was expected to influence post-test outcomes.

To address the nested structure of the data and inter-individual variability, the model included a random intercept for student ID. This accounts for baseline differences in performance between students, capturing individual-level variance that is independent of the experimental conditions.

The complete model was:

$$\begin{aligned}\text{Post-test\_Performance}_i = \beta_0 &+ \beta_1\ \text{Feedback\_Timing}_i(\textit{delayed}(15)/\textit{immediate}) \\ &+ \beta_2\ \text{Answer\_Recall}_i(\textit{recall}/\textit{norecall}) \\ &+ \beta_3(\text{Feedback\_Timing} \times \text{Answer\_Recall})_i \\ &+ \beta_4\text{Pre-test\_Performance}_i \\ &+ \beta_5\text{Nb\_Questions\_Taken}_i \\ &+ \beta_6\text{Students\_Year}_i \\ &+ (1|\text{Student}_i) + \epsilon_i\end{aligned}$$

With i being the identity of the student, and ϵ_i the error term.

Deviations from Pre-registered Analysis Plan In addition to adjustments to inclusion criteria (see Section 2.5), some deviations from the pre-registered analysis plan were implemented to ensure analytical robustness:

- **Design shift to partially between-subjects:** The preregistered design assumed a fully within-subject structure (i.e., each participant contributing data from at least two experimental conditions). In practice, some students contributed data from only one specialty–condition pair. The final analysis was thus conducted as a partially between-subjects design, with a random intercept for student ID to account for repeated measures where applicable.
- **Simplification of random effects structure:** While the preregistration specified a more complex random effects structure including specialty and nested university:student intercepts, the final model retained only a random intercept for student ID. This simplification was necessary due to limited sample size and convergence issues, and was deemed appropriate given that the specialty- and university-level variance was negligible or fully nested within student variance.

3 Results

3.1 Data Filtering

A total of $n = 309$ unique students registered for the experiment. Based on the applied inclusion criteria described in the Methods section (see Section 2.5), participants were excluded at successive stages for incomplete pre-tests, insufficient training engagement, lack of post-test completion, and assignment to the control condition.

First, only students who completed all five pre-test assessments were retained, resulting in $n = 184$ students eligible to proceed to the training phase. Training-related exclusions were then applied in two steps. In the first step, we retained only student–specialty pairs in which students completed at least 15 training questions, yielding 74 students and 154 student–specialty pairs.

In the second step, we retained only students who had met this training threshold in at least one specialty assigned to an experimental condition (i.e., feedback condition $\in \{2, 3, 4, 5\}$), excluding those who had trained exclusively under the control condition. This resulted in a reduced subset of 67 students and 147 student–specialty pairs.

Post-test-related exclusions followed. First, only student–specialty pairs in which students completed at least one isolated-question (IQ) post-test were retained, reducing the dataset to 21 students and 42 student–specialty pairs. Then, to ensure alignment with the experimental design, we retained only those students who had completed post-tests in at least one of the experimental specialties they trained in. This resulted in 20 students and 41 student–specialty pairs.

At this stage, we applied a filter to retain only students in their second year of medical school or higher. This did not engender any additional exclusion. Finally, all remaining student–specialty pairs assigned to the control condition (Condition 1) were removed. This final step ensured that the analytical dataset included only data from the four randomized experimental conditions (Conditions 2–5). The resulting final sample comprised $n = 20$ students and $n = 34$ unique student–specialty pairs.

A detailed overview of the filtering process and sample sizes at each step is provided in the flowchart (see Figure 5.3).

3.2 Descriptive Statistics

The final analytical sample comprised 20 medical students. Of these, 11 students contributed data from a single experimental condition, while 9 students provided data across multiple experimental conditions. Overall, 45% of the participants identified as female, and the sample included students from 13 distinct universities (see Table 5 in the Appendix). The distribution of participants across academic years was as follows: Year 2 ($n = 1$), Year 3 ($n = 3$), Year 4 ($n = 8$), and Year 5 ($n = 8$).

Students reported using a variety of resources to train themselves in their specialty. Among those included in the analysis, 90% reported using the BNE platform, 75% reported using other online platforms, 20% used paper-based annals, and 15% used alternative resources.

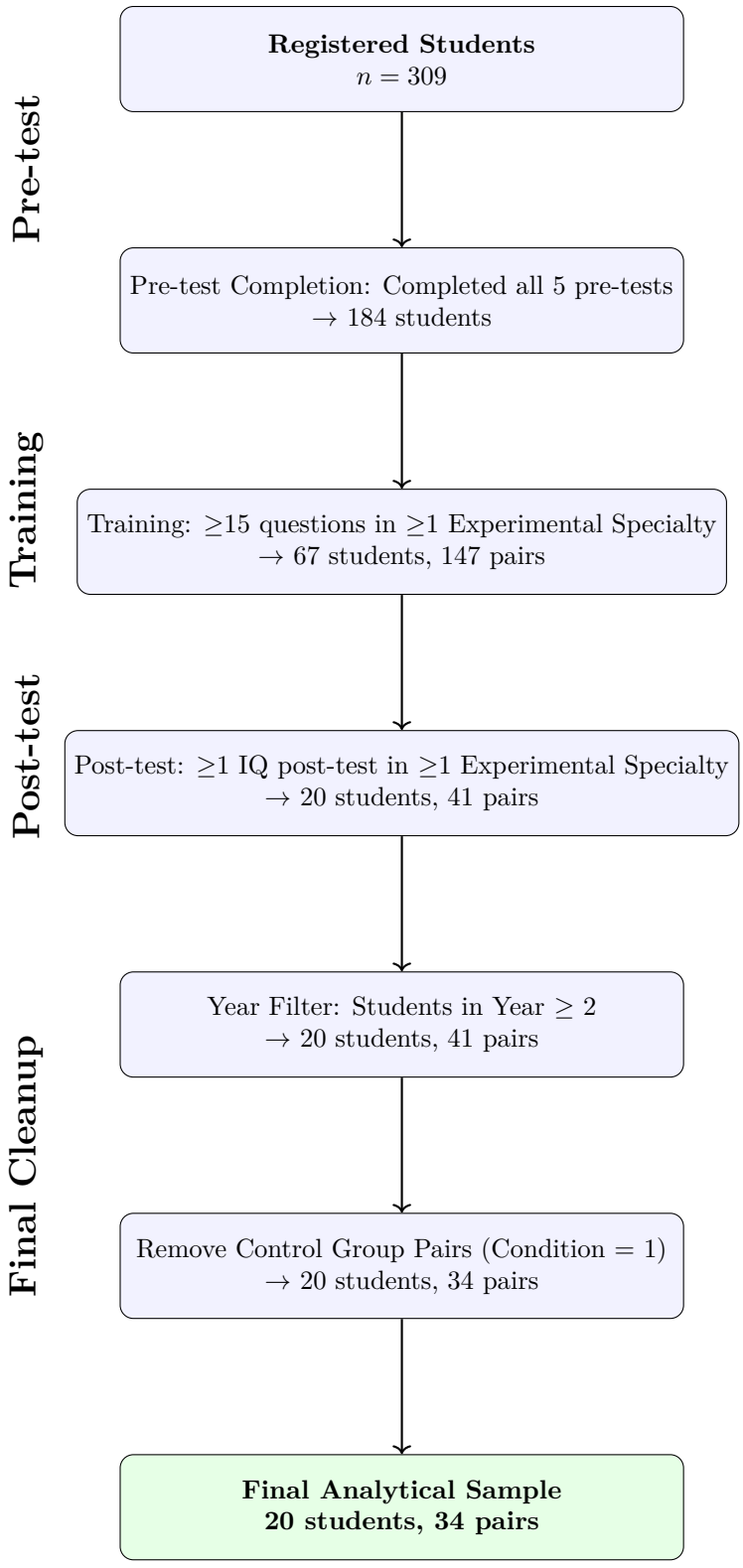


Figure 5.3: Flowchart of participant filtering from registration to final analytical sample.

Table 5.2: Distribution of Specialties Across Feedback Conditions

Specialty	**Feedback Condition**				**Total**
	Delayed + No Recall	Delayed + Recall	Immediate + No Recall	Immediate + Recall	
Ophthalmology	0	0	0	1	1
Gerontology	0	0	1	0	1
Maxillofacial Surgery	0	1	0	0	1
Anesthesiology - Critical Care - Emergency	3	0	0	0	3
Pediatrics	0	1	0	1	2
Rheumatology	1	0	0	1	2
Internal Medicine	0	1	1	0	2
Public Health	0	0	1	0	1
Physical Medicine and Rehabilitation	0	0	0	1	1
Hematology	1	1	1	0	3
Infectious Diseases	0	0	1	0	1
Cardiovascular	3	0	0	2	5
Neurology	0	1	1	0	2
Psychiatry	1	0	0	0	1
Pulmonology	1	0	0	0	1
Hepato-Gastroenterology	0	0	2	0	2
Occupational Medicine	0	0	0	1	1
Dermatology	0	0	2	0	2
Gynecology - Obstetrics	1	0	0	1	2
Total	11	5	10	8	34

The allocation of student–specialty pairs across the four experimental conditions is shown in Table 5.2. Although the experimental algorithm aimed for balanced condition assignment across specialties, the small number of participants created much sparsity in the final dataset.

The mean pre-test score across all included specialties was $M = 0.56$ ($SD = 0.17$). A Shapiro–Wilk test confirmed approximate normality ($W = 0.95$, $p = 0.13$). A one-way ANOVA found no significant differences in pre-test scores between feedback conditions, $F(3, 30) = 0.55$, $p = 0.649$, indicating baseline equivalence across conditions.

During the training phase, students engaged with an average of $M = 1.7$ specialties ($SD = 0.98$), yielding a total of 34 student–specialty pairs. On average, they completed $M = 57.94$ training questions per specialty ($SD = 57.39$, Min $= 15$, Max $= 270$). The distribution of training questions across specialties and conditions is shown in Figure 5.4.

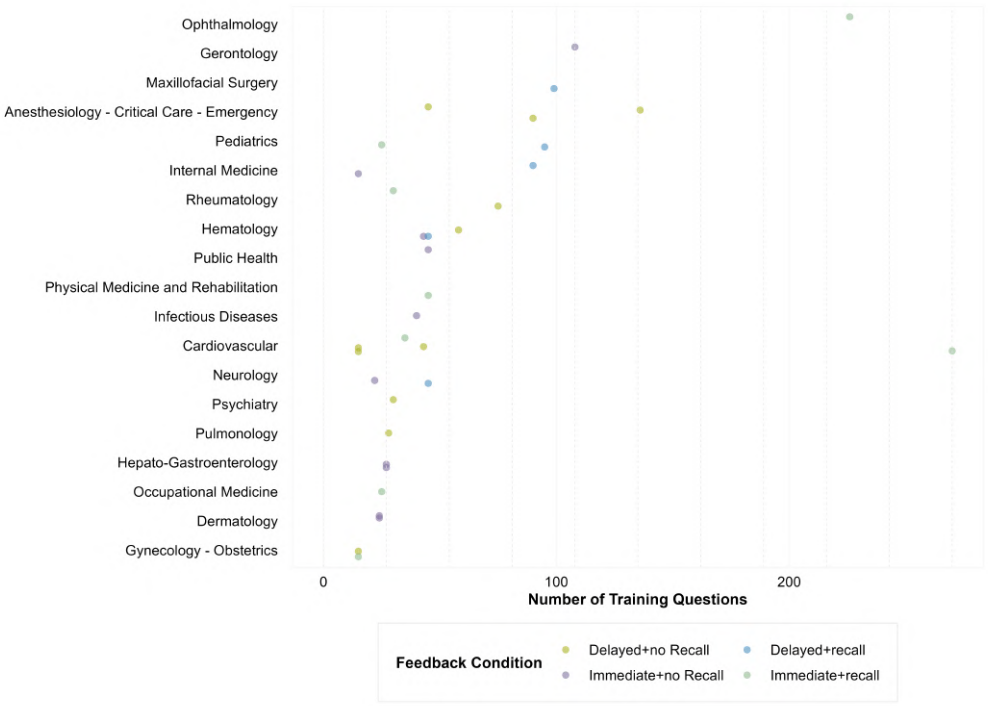


Figure 5.4: Distribution of Training Questions Across Specialties and Feedback Conditions

Table 5.3: Descriptive Statistics by Feedback Condition

	Pre-test		Training		Post-test		Gain	
Feedback condition	M	SD	M	SD	M	SD	M	SD
delayed + no recall	0.61	0.18	0.51	0.13	0.63	0.12	0.02	0.12
delayed + recall	0.52	0.07	0.55	0.13	0.58	0.17	0.06	0.15
immediate + no recall	0.53	0.18	0.50	0.14	0.54	0.12	0.01	0.15
immediate + recall	0.57	0.18	0.57	0.17	0.59	0.16	0.03	0.15

The mean post-test score was $M = 0.59$ ($SD = 0.14$), and the average gain from pre-test to post-test was 0.02 ($SD = 0.14$), indicating minimal improvement overall. Table 5.3 presents condition-wise descriptive statistics for pre-test, training, post-test scores, along with gain scores. Figure 5.5 shows the distribution of pre-test and post-test scores by condition.

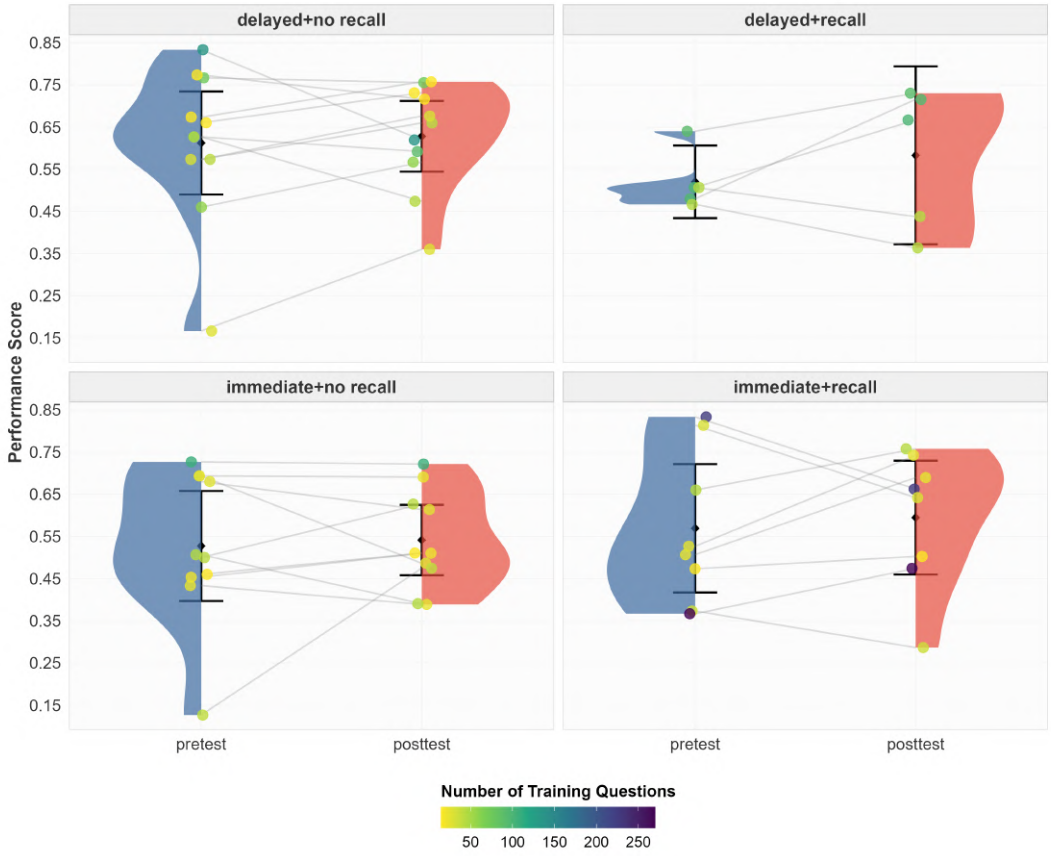


Figure 5.5: Distribution of Pre-test and Post-test Scores by Feedback Condition

3.3 Mixed-Effects Model Results

Pre-test performance was the only significant predictor of post-test performance in this analysis ($\beta = 0.376$, $SE = 0.134$, $t(25.41) = 2.80$, $p = .010$), suggesting that students with higher initial knowledge tended to learn more effectively.

The effect of feedback timing was not statistically significant ($\beta = -0.055$, $SE = 0.046$, $t(21.49) = -1.18$, $p = .251$), suggesting that changing feedback timing alone did not reliably affect outcomes.

Answer recall during feedback had no significant impact on post-test performance ($\beta = -0.0004$, $SE = 0.059$, $t(23.67) = -0.01$, $p = .995$). This suggests that showing students their initial answers did not enhance or impair learning under the current conditions.

The interaction between feedback timing and answer recall was also non-significant ($\beta = 0.051$, $SE = 0.075$, $t(19.83) = 0.69$, $p = .500$). This indicates that the benefit or drawback of immediate versus delayed feedback was not influenced by whether the feedback recalled the initial responses.

Figure 5.6 presents the adjusted post-test performance scores predicted by the mixed-effects model across feedback timing and initial answer recall conditions. The plot illustrates the estimated marginal means with 95% confidence intervals, highlighting the lack of interaction or

Table 5.4: Linear Mixed-Effects Model Results Predicting Posttest Performance

Fixed Effects					
Predictor	**Estimate**	**SE**	**df**	***t*-value**	***p*-value**
Intercept	0.371	0.131	23.25	2.83	0.010*
Feedback Timing: Immediate	-0.055	0.046	21.49	-1.18	0.251
Answer Recall: Recall	-0.001	0.059	23.67	-0.01	0.995
Pretest Performance	0.376	0.134	25.41	2.80	0.010*
Number of Training Questions	-0.0002	0.0004	26.59	-0.50	0.619
Students' Year	0.009	0.028	18.91	0.33	0.747
Immediate × Recall	0.051	0.075	19.83	0.69	0.500
Random Effects					
Student ID (Intercept)		$\sigma^2 = 0.004$, $\sigma = 0.064$			
Residual		$\sigma^2 = 0.009$, $\sigma = 0.096$			
Model Fit					
AIC / BIC		-0.81 / 12.93			
Log Likelihood		9.40			
Observations		34			
Participants (Student ID)		20			
Variance Inflation Factors		Range: $1.17 - 3.14$			

Note: $^*p < .05$. SE = Standard Error. Feedback Timing reference level: Delayed. Answer Recall reference level: No Recall.

Figure 5.6: Adjusted post-test performance across feedback timing and initial answer recall conditions. Error bars represent 95% confidence intervals around the estimated marginal means.

main effects of the two factors.

The number of training questions completed ($\beta = -0.0002$, $p = .619$) and student year ($\beta = 0.009$, $p = .747$) were not significant predictors. This implies that variation in training quantity and academic level did not influence post-test outcomes.

The random intercept for student ID explained 30.4% of the total variance ($\sigma^2 = 0.00405$), indicating moderate individual differences in learning outcomes. Residual variance was estimated at $\sigma^2 = 0.00928$, representing the remaining within-student variation not accounted for by the random effect.

To assess multicollinearity among predictors, variance inflation factors (VIFs) were examined. All VIF values were below 3.2, with the highest being for the interaction term (VIF = 3.14). These values indicate no serious multicollinearity concerns and suggest that the model estimates are stable and interpretable (Kim, 2019; Sheather, 2009).

4 Discussion

This study set out to examine the effects of feedback timing (immediate vs. delayed), initial answer recall, and their interaction on post-test performance within the BNE digital education platform. Although the experimental design was developed to test theoretically motivated hypotheses, the results did not yield statistically significant main effects or interactions. Given the size and structure of the final analytical sample, these null findings should not be interpreted as conclusive evidence against the effects of feedback timing or answer recall, but rather as reflective of the paucity of data.

A central limitation of this study was the small sample size, which ultimately constrained the analyses and required several deviations from the preregistered plan. This limitation stemmed primarily from the unexpectedly long timeline required to develop, test, and integrate the experimental module within the BNE platform. The full implementation process spanned over two years, which delayed the experiment to the final phase of the thesis timeline. As a result, there was insufficient time to keep the study open long enough to reach the target sample. Although strong engagement was expected, particularly given that the standard BNE module is widely used and was reported by participants as a primary training resource, participation in the experimental module fell short. Although the major benefit of running the experiment on this online platform was initially thought to reside in the access to a large population of students (about 8800 per year), the communication channels available to us did not suffice to generate much participation. Additional constraints, such as the academic calendar and competing demands on students' time, may have further contributed to the low completion rates.

As a result, only 20 students and 34 student–specialty pairs met the final inclusion criteria, which is substantially fewer than the 139 participants estimated by the original power analysis to be required for detecting a small-to-medium effect size with 80% power in a within-subjects design (and this assumed that all students would train in 5 specialties). The limited sample size imposed a shift to a partially between-subjects design, in order to rescue more participants, but at the cost of increasing between-condition variability. In parallel, lower-than-expected

engagement during the training and post-test phases required relaxation of the preregistered data-filtering thresholds to retain a sufficient number of analyzable cases.

The simplification of the random effects structure was also driven by the limited sample size and related analytical constraints. While the preregistered model specified random intercepts for specialty and nested university-student structures, these components were omitted in the final model due to convergence issues, insufficient observations per grouping level, and the overall sparsity of the data. Such modeling simplifications were necessary to ensure estimability, but they likely reduced the model's ability to account for contextual variation in feedback reception. This is particularly relevant in education research, where students operate within hierarchical systems defined by institutional curricula and content-specific pedagogies. As emphasized by Leppink (2015), multilevel modeling is essential in educational research for disentangling individual learning effects from broader contextual influences, thereby improving both the validity and precision of statistical inferences. Thus, the exclusion of these contextual levels in our final model likely contributed to reduced model precision.

Furthermore, the decision to reduce the minimum training threshold from 60 to 15 questions per student–specialty pair was also driven by the limited sample size and lower-than-expected student engagement. While this adjustment was necessary to retain a workable analytical sample, it also meant that some participants may have received insufficient exposure to the feedback manipulation for any learning effects to meaningfully emerge. Similarly, the post-test inclusion criterion, which required only a single completed isolated-question (IQ) test per specialty, was adopted to maximize data retention but may have decreased the robustness of the estimation oflearning gains, particularly in such a complex and higher-order educational domain.

Despite these limitations, the study design, if implemented as intended, has strong potential for both theoretical and practical contributions. With sufficient statistical power and balanced within-subject exposure, the preregistered analysis could offer meaningful insights into how feedback timing and initial answer recall influence learning outcomes, especially in complex domains like clinical reasoning and specialty-based education.

Theoretically, the study could advance our understanding of the interference perseveration theory by isolating the effect of feedback timing from answer recall in higher-order learning. Practically, a fully powered implementation could have informed improvements to the BNE platform. The findings could guide the optimization of feedback strategies to support long-term retention and application of knowledge in medical training. Although current findings remain preliminary, the study provides a strong foundation for such future applications.

Importantly, data collection is still ongoing, and additional participation over the coming months may enable a more robust and fully powered analysis. The experimental module has been reopened from the pre-test phase, allowing new students to enroll while enabling ongoing participants to complete training and post-tests in their assigned specialties. With a larger dataset, it may be possible to reinstate the original within-subjects design, and implement a more complex random effects structure.

To support these goals, further efforts may be needed to increase student engagement. While weekly reminder emails have already been sent throughout the study, these may not be sufficient

to maintain motivation over time. For the upcoming data collection phase, additional strategies could include offering performance feedback reports earlier, providing small participation incentives, or collaborating more closely with faculty to promote participation through official course channels. Such measures could help increase both completion rates and the depth of student engagement with the experimental tasks.

In summary, the current results should be interpreted primarily through the lens of methodological limitations. The null findings do not permit strong conclusions about the educational efficacy of delayed feedback or answer recall in digital medical learning. However, with continued data collection and methodological refinement, the design has the potential to address meaningful questions about feedback in technology-enhanced education.

5 Appendix

0.1 University Distribution

Table 5: Distribution of Universities

University Name	Nb Students
Université de Poitiers	5
Université de Lille	3
Université de Nantes	2
Sorbonne Université	1
Université Clermont-Auvergne	1
Université Paris Cité	1
Université Paris-Saclay (XI)	1
Université d'Angers	1
Université de La Réunion	1
Université de Lyon-I	1
Université de Montpellier	1
Université de Rennes	1
Université de Toulouse-III	1

Chapter 6

General Discussion

The objective of this chapter is to synthesize the main findings of the dissertation and discuss their implications. It begins by summarizing the results of each empirical study in relation to the overarching research questions. Then it offers an integrated perspective that connects learner modeling, training difficulty, and feedback timing as interdependent components of effective instructional design. Finally, it outlines the broader theoretical, methodological, and practical contributions of the dissertation, while acknowledging its limitations.

Contents

1 Summary and Contributions of the Studies

1.1 Main Results of Chapter 2 (Learning Analytics Study)

Could the Elo rating system be adapted to a particularly challenging real-life scenario, such as the BNE platform, to provide real-time estimations?

To address this first research question of my doctoral study, we extended the Elo rating system to accommodate the unique challenges posed by the BNE dataset. These challenges include a large-scale and highly structured medical knowledge corpus, substantial overlap between knowledge components, sparse user-question interactions, and a diverse student population. We then compared the accuracy of our adapted Elo rating system against the widely accepted logistic regression model, which does not provide real-time estimations.

Our findings indicate that, despite the complexities of the dataset, the multi-concept Elo rating system performed comparably to logistic regression in predicting final exam outcomes, achieving an of 0.812 and an accuracy of 73.7%, against = 0.811 and accuracy = 73.6% for logistic regression. **This suggests that the Elo rating system can dynamically adjust to student performance and question difficulty in real-time, making it an online option for adaptive learning platforms, even in complex environments such as the BNE**.

The multi-concept Elo rating systems's ability to estimate student proficiency across multiple medical specialties was particularly beneficial in this dataset, where questions often required knowledge from multiple domains. Given its strong predictive accuracy, the adapted Elo rating system presents a promising alternative for real-time student assessment in large-scale educational platforms by offering both computational efficiency and transparency while maintaining robust performance.

These findings motivated further exploration of the adapted Elo rating system within the BNE, particularly to investigate optimal training difficulty. By leveraging its ability to dynamically estimate both student ability and item difficulty, we used it in the subsequent project to estimate real-time relative difficulty to measure how challenging a question is for a student, given their proficiency at answering. This initial analysis thus serves as the foundation for the broader scope of this doctoral research.

How can large initial errors in the Elo rating system be mitigated?

Although the adapted Elo rating system demonstrated good predictive accuracy within the BNE platform, the standard initialization method, which assigns all initial ratings to zero, often results in high uncertainty during early iterations. To address this limitation, we explored the potential benefits of initializing the Elo rating system with estimates derived from logistic regression applied to data from preceding years. This approach aimed to enhance early-stage prediction accuracy by leveraging historical performance data.

Our results indicate that this initialization method led to an initial improvement in (+0.016) and a reduction in (−0.008) during the first 30 days of training, compared to the Elo rating

system initialized at zero. **These findings suggest that historical performance data can help reduce early estimation errors and accelerate model convergence**. However, as the system accumulated more student interaction data, the predictive advantage of historical initialization decreased, resulting in a final improvement of only $+0.002$ and -0.002 by the end of the training period.

While the long-term impact on overall predictive performance remains marginal, this approach offers practical advantages in educational settings where early predictions are crucial for guiding adaptive learning interventions. Thus, initializing Elo ratings with logistic regression outputs may serve as a valuable strategy for improving real-time learning analytics in large-scale educational platforms.

1.2 Main Results of Chapter 3 (Quasi-experimental Study)

Can an optimal level of training difficulty be determined for medical students using the BNE digital learning platform?

Given the platform's randomized question assignment mechanism, we designed a quasi-experimental study leveraging an adapted multi-concept Elo rating system to dynamically estimate both student proficiency and question difficulty. Within this framework, we computed mean relative difficulty—the average difference between the difficulty of encountered questions and a student's ability at the time of answering—over the course of their training. We then applied a linear mixed-effects model to assess the impact of this relative difficulty on final exam performance, controlling for students' initial ability, training intensity, and specialty-specific factors.

Our findings revealed a significant **quadratic relationship** between training difficulty and final exam outcomes, confirming the Inverted U-shaped Hypothesis. **This result provides strong evidence for the existence of an optimal training difficulty level for medical students using the BNE digital learning platform, which corresponds to the apex of the performance curve, at which learning outcomes are maximized**. These findings underscore the importance of dynamically adjusting training difficulty within adaptive educational systems to ensure that students are consistently challenged at an optimal level, thereby enhancing learning efficiency and long-term retention.

What is the optimal level of training difficulty in this specific context, and does it vary between medical specialties?

Having established the existence of an optimal training difficulty level, we examined its specific value and whether it varied across medical specialties. Our analysis revealed that optimal difficulty is not a fixed threshold but varies based on both the medical specialty and student ability level. **On average, students achieved the highest final exam performance when exposed to questions with a relative difficulty between -1.55 and -0.53, corresponding to a success probability of 64% to 83% for multiple-answer questions in the BNE**.

However, **this optimal range differed across medical specialties, reflecting differences in domain complexity and other domain specific factors**. Additionally, student ability played a crucial role, with higher-ability students demonstrating greater sensitivity to training difficulty, while lower-ability students were less affected by variations in question difficulty. This pattern, discussed in detail in Chapter 3, may in part be explained by the tendency of lower-ability students to respond in a relatively random or inconsistent manner, regardless of question difficulty, thereby reducing their apparent sensitivity to this factor.

Together with the findings from Chapter 2, our results suggest that tailoring training difficulty to individual learning trajectories and subject-specific factors should provide an optimal challenge level and ultimately enhance learning efficiency. Moreover, achieving this level of personalization is feasible through a well-adapted and rigorously tested learner model which are specifically designed to address the unique requirements of the learning platform.

1.3 Main Results of Chapter 4 (Meta-analysis)

What is the difference in the effect of immediate versus delayed feedback on learning?

To investigate the impact of feedback timing on learning outcomes, we conducted a meta-analysis comparing immediate versus delayed feedback within computerized learning environments. This involved quantifying the effect size of learning outcomes across 51 studies published since 1988. **Our analysis revealed no statistically significant difference in learning outcomes between immediate and delayed feedback conditions** ($g = 0.03$, 95% CI [−0.07, 0.14], $p = 0.518$). In other words, the timing of feedback presentation, on average, did not consistently influence learning performance.

However, substantial heterogeneity was observed across studies ($I^2 = 82\%$, $\tau^2 = 0.12$), suggesting that the effect of feedback timing may be moderated by other factors. This significant variability prompted us to conduct a moderator analysis to explore potential factors that might influence the effect of feedback timing, leading to our second research question.

How do varying definitions of "immediate" and "delayed" feedback influence the reported effectiveness of these feedback timings on learning outcomes?

Our meta-analysis examined how varying definitions of "immediate" and "delayed" feedback influence the reported effectiveness of these feedback timings on learning outcomes by considering how delay was operationalized across studies as a moderator. We identified two key sources of variation in definitions. First, studies differed in how they measured delay—whether in absolute time (e.g., seconds), the number of intervening items, or a combination of both. Second, the magnitude of the delay difference between immediate and delayed feedback varied across studies.

Our results indicated that the choice of measurement unit (time vs. items) did not significantly impact overall learning outcomes. However, when analyzing the extent of the delay difference, we found that for studies using item-based delays, greater delays were associated

with an increasingly positive effect of delayed feedback on performance. This effect was not observed when delays were measured in seconds or when item-based delays were converted into time units. **These findings suggest that the way immediate and delayed feedback are defined does not systematically influence their reported effectiveness, which contradicts expectations in the literature** (Kulik et al., 1988; Mory, 2013; M. Xu et al., 2023).

This moderation analysis was limited by inconsistencies in how delay units and delay differences were reported across studies, preventing precise coding. These findings highlight the need for greater consistency in defining feedback timing in future research to facilitate more accurate comparisons and interpretations.

What other factors moderate the effects of immediate versus delayed feedback on learning outcomes?

To address this question, we conducted moderator analyses based on study and feedback characteristics.

The results indicated that the type of feedback influenced the effect of feedback timing—studies using try-again feedback showed a significant advantage for delayed feedback, whereas studies using elaborated and simple feedback types did not. Educational level also played a role, with delayed feedback benefiting learners in primary and secondary education, while no significant effect was observed for tertiary and adult learners. Additionally, learning domain moderated feedback timing effects, with delayed feedback being more effective for text-based learning tasks, such as reading comprehension and text memorization, whereas immediate feedback showed advantages in STEM, language learning, and cognitive skill tasks. The nature of the post-test task further influenced outcomes, with memory retrieval tasks favoring delayed feedback and knowledge application tasks showing a slight, though non-significant, advantage for immediate feedback. Finally, response time constraints moderated feedback effects—when learners had unlimited time, delayed feedback was more effective, whereas, under time-limited conditions, no clear advantage was observed.

However, when considering these moderators simultaneously in a multiple meta-regression, none remained statistically significant, suggesting that these factors may be interrelated.

1.4 Main Results of Chapter 5 (Experimental Study)

How do feedback timing (immediate vs. delayed) and recall of initial responses interact to influence learning outcomes in higher-level medical training?

To address this question, we conducted a randomized controlled experiment within the BNE Expérimentale module, embedded in the UNESS digital learning platform. Medical students were assigned to one of four experimental conditions in a 2×2 factorial design that crossed feedback timing (immediate vs. delayed) with initial answer recall (recall vs. no recall) during

training on multiple-choice questions.

The analysis did not reveal statistically significant main effects or interactions of feedback timing and initial response recall on post-test performance. While these null results may suggest limited impact of the manipulations under the current conditions, they are likely attributable to several contextual constraints. In particular, the small sample size, modest participant engagement, and limited exposure to the feedback conditions reduced the statistical power to detect potential effects.

Nevertheless, the study demonstrates the feasibility of embedding controlled experiments within a large-scale digital education platform. Importantly, data collection remains ongoing, and further participation is expected to support more robust analyses. These future analyses may clarify whether, and under what conditions, feedback timing and response recall influence learning outcomes in complex, higher-order domains such as clinical medical training.

2 General Synthesis

Taken together, the findings from these four studies converge on a central insight: optimizing learning in complex digital educational environments requires an integrated, personalized approach that bridges cognitive science with scalable algorithmic modeling.

First, the thesis demonstrates the feasibility and advantages of real-time learner modeling using an adapted rating system in large-scale, complex educational settings. Second, building on the learner model developed in the first study, the research shows that training difficulty must be personalized as its impact depends on both the learner's profile and the nature of the learning task. Third, the meta-analysis highlights the importance of tailoring feedback strategies to the specific demands of the learning domain, the learner's educational level, and other contextual factors. While the experimental study did not yield conclusive evidence on the effects of feedback timing or answer recall, it illustrates the practical challenges of implementing embedded experiments in authentic educational environments and offers preliminary insights to guide future research in this area.

The studies are presented in the order in which they were conducted, based on data availability and their interdependencies. However, rather than treating learner modeling, training difficulty, and feedback as separate dimensions, they should be viewed as interconnected components of an effective learning system. Ultimately, this research advocates for an adaptive approach that integrates feedback delivery, difficulty calibration, and personalization in a balanced and coordinated manner to optimize learning outcomes in digital environments.

At the heart of this integration is the concept of real-time adaptivity. The adapted multi-concept Elo rating system developed in this work serves not only as a predictive tool but also demonstrates the feasibility of moving beyond static, one-size-fits-all educational models toward systems that can respond to learners' evolving profiles. While this model was not deployed in an adaptive loop, the findings suggest it has strong potential for driving real-time personalization. Importantly, it was validated in a national-scale, high-stakes medical training context. This real-world validation reveals that, even in the face of extreme data sparsity, overlapping content

domains, and heterogeneity in learner engagement, it is possible to obtain robust, interpretable, and responsive estimates of both student proficiency and item difficulty.

This capacity for real-time estimation becomes especially critical when considering the findings on training difficulty, as a central research challenge lies in accurately measuring both task difficulty and learner skill. Previous research has emphasized that item difficulty should be understood as a relative construct which depends on the learner's ability at the time of engagement, rather than as an inherent or absolute property (Abuhamdeh and Csikszentmihalyi, 2012; Gallego et al., 2018; R. C. Wilson et al., 2019). Building on this perspective, our study employed the adapted Elo rating system to dynamically estimate relative difficulty throughout the training process. This approach enabled us to identify an optimal success probability range for five-option multiple-choice questions. This range not only aligns with prior findings suggesting that relatively high success rates promote long-term retention but also resonates with literature from game-based learning, where such difficulty levels are shown to foster intrinsic motivation by sustaining learners in a state of flow (R. C. Wilson et al., 2019; Ninaus et al., 2017; Perttula et al., 2017).

Taken together, these findings suggest that adaptivity in digital learning environments must operate at two levels. First, at the algorithmic level, learner models like the multi-concept Elo rating system must continuously update estimates of student ability and item difficulty in response to data. Second, at the instructional level, systems must leverage these real-time estimates to personalize the difficulty of content.

Feedback, another cornerstone of learning theory, also emerges in this thesis as a highly context-sensitive mechanism. The meta-analysis reveals that the effects of feedback timing are not uniformly distributed but vary systematically with study-level characteristics. Crucially, many of these characteristics—such as educational level, learning domain, task demands, and response time constraints—were found to covary with publication year. These temporal patterns reflect broader shifts in instructional design and research focus, suggesting that changes in how feedback is implemented and studied over time have shaped the observed effects. Rather than supporting a fixed advantage for either immediate or delayed feedback, the findings highlight that its effectiveness is contingent on the evolving context in which learning takes place.

The experimental study, while limited in scope and statistical power, complements this conclusion by illustrating the difficulty of isolating feedback effects in real-world, high-variability learning environments. Although it did not yield reliable effects of feedback timing or answer recall, it helped identify practical implementation constraints, such as participation drop-off and variability in engagement, that must be addressed in future embedded studies. As such, its contribution lies in informing the design and feasibility of more robust experiments within authentic digital education systems.

3 Contributions

As highlighted by Koedinger, Booth, et al. (2013) and Schneider et al. (2017), optimal learning outcomes are most likely to arise when the learning systems are both theoretically principled

and empirically validated. This thesis adheres to that vision by integrating computational techniques with insights from cognitive science and by testing these systems at scale within authentic educational contexts.

Building on this foundation, the thesis makes contributions to the fields of educational technology, learning analytics, and cognitive psychology by advancing our understanding of how digital learning systems can be effectively designed, implemented, and validated in real-world settings.

One of the central methodological contributions of this dissertation is the adaptation and validation of a scalable, multi-concept Elo rating system. By adapting the Elo rating system to handle the structural and statistical characteristics of the BNE medical training platform, the thesis provides a methodological advancement over traditional models such as logistic regression, particularly in its ability to operate in real-world, high-stakes contexts.

This dissertation also makes important theoretical contributions. It empirically validates and extends the Inverted-U shape hypothesis and shows that optimal training difficulty is not a static threshold but varies with learner ability and domain complexity. Furthermore, this dissertation refines our understanding of feedback timing by integrating findings from a large-scale meta-analysis.

Practically, this thesis offers a framework for designing adaptive educational systems that dynamically adjust the difficulty of learning materials in real-time and tailor the timing of feedback based on domain, learner, and contextual characteristics. The findings highlight the importance of personalization which is granular and sensitive to the intersections of learner, task, and context. These principles are particularly relevant for intelligent tutoring systems that aim to keep learners within their optimal zone of proximal development (L. S. Vygotsky et al., 1978), provide appropriately timed challenges (E. L. Bjork et al., 2011), and support cognitive processes such as retrieval and error correction.

Additionally, this work contributes empirically by leveraging a large medical training dataset that includes extensive learner-item interactions and high-stakes assessment outcomes. This scale of validation strengthens the generalizability of the findings and demonstrates the feasibility of integrating learning science theory into operational systems used by diverse learner populations.

Together, these contributions bridge the gap between theory and application in adaptive learning. They offer a comprehensive framework—both conceptual and computational—for creating intelligent learning technologies that are empirically grounded, cognitively informed, and scalable in practice.

4 General Limitations

While this dissertation contributes novel insights into optimizing adaptive digital learning systems within the context of medical education, it is important to recognize several broad limitations that influence the interpretation and generalizability of the findings. In addition to study-specific constraints described in individual chapters, this section outlines three major cat-

egories of limitations relevant to the thesis as a whole: (1) external validity and generalizability, (2) platform-specific constraints, (3) measurement and data limitations, and (4) challenges related to causal inference.

4.1 External Validity and Generalizability

A key limitation of this dissertation lies in the specificity of its empirical context. Three of the empirical studies—the adaptation of the Elo rating system, the quasi-experimental analysis of training difficulty, and the experimental study on feedback timing—primarily draw on data from the BNE digital learning platform, which is predominantly used by French medical students. However, the unique structure and characteristics of BNE limit the generalizability of the findings.

For example, the adapted Elo rating system presented in Chapter 2 was specifically tailored to address the multi-specialty organization and high sparsity of the BNE question bank. While the model demonstrated strong predictive performance in this context, its applicability to platforms involving open-ended tasks, alternative instructional formats, or different domain requirements remains untested. Similarly, the training difficulty estimates discussed in Chapter 3 were based largely on multiple-choice questions with five response options. These findings may not directly transfer to contexts that emphasize clinical reasoning, open-ended problem-solving, or collaborative learning. Moreover, although this dissertation focuses on medical education, prior research has shown that even adaptive learning systems—which dynamically adjust content to individual learners—are heavily influenced by domain-specific instructional design features (L. Chen et al., 2020).

Therefore, generalizing the findings to other educational domains, disciplines, or learning materials or environments should be approached with caution, since learning analytics conducted within a single platform often reflects the unique instructional design, assessment methods, and learner characteristics of that environment (Gašević et al., 2015).

4.2 Platform-Specific Limitations

Beyond these external validity limitations, the BNE platform itself presents several structural and operational constraints. First, the BNE is a non-adaptive platform and does not dynamically tailor content to learner performance. In contrast, many advanced digital learning environments employ real-time dynamic item selection based on learners' prior responses (e.g., J. R. Anderson et al. (1995) and Ma et al. (2014)). As such, the findings regarding optimal difficulty and feedback timing are limited to non-personalized environments and may differ in adaptive learning systems where content presentation evolves responsively during the learning process.

Another important limitation concerns the challenges posed by the BNE platform in reliably estimating student ability over time. Because participation is voluntary, students engage with the platform to varying degrees and with irregular frequency. This results in uneven data density and potential self-selection bias, where more motivated or higher-performing students might be overrepresented in the dataset (Wise, 2014). Moreover, this irregular engagement

presents significant challenges for learner modeling—particularly for students who are inactive for long periods and later return. Elo-based models, which update learner and item estimates incrementally based on recent interactions, are sensitive to such temporal discontinuities. As shown by Vermeiren et al. (2025), prolonged learner inactivity can contribute to rating drift in item parameters, whereby repeated learner progress gradually deflates item difficulty estimates. When returning students re-engage, they may encounter items that appear easier than they actually are, leading to inflated difficulty and misalignment with their current ability.

Additionally, student learning is not confined to the BNE platform. Many learners rely on external resources, such as textbooks, lectures, peer discussions, or clinical practice, none of which are captured in the platform data. These off-platform activities introduce unobserved variability that further complicates the interpretation of platform-derived ability estimates.

More broadly, the BNE platform ultimately proved less well-suited to the experimental goals initially envisioned for this project. While it offered the promise of testing learning interventions at scale, in practice, it was difficult to implement experimental versions of the platform. Technical and logistical barriers made it challenging to manipulate content delivery or integrate real-time adaptivity into the system. Even a relatively simple feedback experiment required extensive coordination, and another planned experiment on training difficulty could not be implemented due to time constraints. Recruitment also fell short of expectations, and participation levels varied widely across students and time, further increasing noise in both the observational and experimental data. In retrospect, simpler and more flexible platforms combined with more tightly controlled curricula may offer a more practical environment for deploying adaptive interventions and conducting iterative experimentation. While the BNE remains a valuable source for observational research and offers a unique opportunity to test learning optimization strategies in complex, higher-level medical education, translating findings into platform-level changes proved significantly more complex than initially anticipated. This reflects an important lesson for future research aiming to bridge educational theory with real-world system design.

4.3 Measurement and Data Limitations

Several measurement-related limitations constrain the interpretation of this dissertation's findings. First, the primary outcome measure across all empirical studies was post-test performance on (). While s provide an objective, scalable, and time-efficient assessment format, they do not adequately capture deeper cognitive processes such as critical thinking, clinical reasoning, or complex problem-solving skills (Case et al., 1998; Schuwirth et al., 2011). This limits the extent to which observed learning outcomes can be generalized to more complex or applied educational competencies.

Second, the temporal granularity of the data imposed significant restrictions. On the BNE platform, timestamps were recorded only at the session level rather than at the level of individual student–question interactions. This limitation made it impossible to extract key indicators such as time-on-task or response latency—fine-grained behavioral measures often used to infer engagement, self-regulated learning strategies, attention levels, or cognitive processing effort (Naumann, 2019; Seufert, 2020; Loon et al., 2023). The absence of such data weakens the

interpretability of student interaction patterns and learning behaviors.

Third, the dataset's structure introduced several domain-specific ambiguities. The user–question interaction matrix was highly sparse—each student had attempted only a small fraction of the total question bank. While this level of sparsity is typical in large-scale educational datasets (e.g., Junyi, ASSISTments, EDNet), it poses challenges for statistical modeling and can lead to biased or unstable parameter estimates (Greenland et al., 2016).

Additionally, many questions were assigned to multiple medical specialties, but these tags were unweighted, making it unclear how much each specialty contributed to a question's content. This required assumptions about content attribution that may have introduced noise into specialty-based analyses. Furthermore, a significant number of questions were excluded entirely due to missing specialty tags, further limiting the dataset.

Furthermore, while students were free to choose the specialties they wanted to focus on during training, the dataset did not capture their intended specialty preferences. Instead, specialty information was derived from instructor-assigned tags at the question level. Because many questions were associated with overlapping specialties and the system did not register students' declared training intentions, it was not possible to reliably infer which specialty a student was targeting at any given time. As a result, it was difficult to assess learners' self-regulated strategies, such as topic selection, training intensity within a specialty, or spacing between practice sessions—factors known to be critical for effective learning.

Taken together, these limitations significantly hindered the operationalization of engagement and self-regulated learning within this research. Prior studies have shown that such variables can meaningfully mediate the effectiveness of adaptive learning interventions (Plooy et al., 2024; Papoušek, Stanislav, et al., 2016b; Nkhoma et al., 2014). The absence of these behavioral and motivational indicators reduces the explanatory power of the models used and limits the ability to draw conclusions about the underlying learning mechanisms driving observed outcomes (Gasevic et al., 2017).

4.4 Challenges to Causal Inference

Although one of the studies (Chapter 5) employed a randomized controlled trial, the majority of analyses in this dissertation relied on quasi-experimental or observational data. For instance, in the analysis of optimal training difficulty (Chapter 3), question exposure varied naturally. While covariates such as engagement were statistically controlled by adding the number of training questions as a covariate, unobserved confounders may still have influenced outcomes. One notable limitation is that prior knowledge at the specialty level was not directly measured or controlled. Instead, final student ability at the end of training was included as a covariate, which reflects post-exposure performance rather than pre-existing proficiency. A fully experimental design could have addressed this more precisely by incorporating a specialty-level pretest, as implemented in our experimental study. Moreover, the quasi-experimental design relied on students' naturally encountered mean difficulty levels within each specialty, based on the random draw of questions, which does not guarantee balanced exposure. A within-subject experimental design—randomly assigning difficulty levels across specialties for each student—would provide

stronger causal identification. Finally, variables such as motivation, effort, or off-platform learning could not be observed or reliably estimated. While these are difficult to control even in experimental settings, they likely contributed to residual confounding. These limitations reduce the strength of causal claims that can be drawn from the quasi-experimental findings (Gopalan et al., 2020).

In addition to these design constraints, online learning environments inherently lack strict experimental control. For example, it is difficult to ensure that participants are fully engaged or adhering to protocols. Participants may have been multitasking, consulting outside materials, or experiencing distractions during training, all of which may have affected the outcomes. It was impossible to control for these events.

Moreover, as demonstrated in Chapter 5, even randomized experiments conducted in real-world educational platforms face substantial challenges. Despite the rigor of the design, the implementation was constrained by low participation, limited engagement, and practical limitations on exposure to the experimental conditions. These factors reduced statistical power and the interpretability of the results, highlighting that randomization alone does not guarantee strong causal inference when real-world variability and logistical barriers limit treatment delivery and data quality.

Thus, isolating the effects of specific instructional variables (e.g., feedback timing) is challenging in multifactorial environments where learners' behaviors, prior experiences, and contexts interact in complex ways. Without fine-grained control or data triangulation, attributing performance outcomes solely to manipulated variables remains difficult, even under experimental conditions.

In sum, while the findings of this dissertation are grounded in ecologically valid, large-scale data and complemented by rigorous analytic approaches, several limitations remain. These include external validity and generalizability limitations, contextual and platform-specific constraints, measurement and data challenges, and the difficulty of establishing strong causal inferences. Recognizing these limitations not only informs the appropriate interpretation of results but also highlights directions for future research, including the integration of fine-grained process data, broader cognitive assessments, and experimental replication in adaptive, diverse learning environments.

Conclusion and Perspectives

To conclude, this dissertation has provided new insights into the cognitive and contextual factors that shape learning in digital educational environments, with a particular focus on learner modeling, training difficulty, and feedback delivery strategies. The results carry practical implications for researchers, instructional designers, and education professionals, and offer guidance for the development of future digital learning systems.

The collection of studies presented here underscores the importance of integrating algorithmic precision with cognitive theory to support real-time personalization. By aligning learning analytics with principles from educational psychology and cognitive science, this research lays the groundwork for next-generation platforms that are not only scalable and data-driven, but also cognitively informed. Future work should continue testing these principles through experimental interventions in authentic learning contexts, thereby advancing both theory and practice in the field of digital learning.

Equally important is the need for closer collaboration between researchers, developers, and educators to ensure that adaptive systems reflect the complexity of human learning and the diversity of learners. As educational technology continues to evolve, the central challenge remains: how can data and algorithms be used not merely to optimize performance, but to foster meaningful, durable, and equitable learning?

This work represents one step in that direction. The insights it offers are not intended as end points, but as building blocks for the next generation of adaptive and human-centered learning environments.

References

Abdelrahman, Ghodai, Qing Wang, and Bernardo Nunes (2023). "Knowledge tracing: A survey". *ACM Computing Surveys* 55.11, pp. 1–37 (cit. on p. 37).

Abdi, Solmaz, Hassan Khosravi, and Shazia Sadiq (2021). "Modelling learners in adaptive educational systems: A multivariate glicko-based approach". *Lak21: 11th international learning analytics and knowledge conference*, pp. 497–503 (cit. on pp. 22, 53, 66).

Abdi, Solmaz, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic (2019). "A multivariate Elo-based learner model for adaptive educational systems". *arXiv preprint arXiv:1910.12581* (cit. on pp. 21, 22, 37, 44, 45, 53).

Abuhamdeh, Sami and Mihaly Csikszentmihalyi (2012). "The Importance of Challenge for the Enjoyment of Intrinsically Motivated, Goal-Directed Activities". *Personality and Social Psychology Bulletin* 38.3, pp. 317–330 (cit. on pp. 22, 24, 25, 58, 76, 78, 173).

Abuhamdeh, Sami, Mihaly Csikszentmihalyi, and Baland Jalal (2015). "Enjoying the possibility of defeat: Outcome uncertainty, suspense, and intrinsic motivation". *Motivation and Emotion* 39, pp. 1–10 (cit. on p. 78).

Ahn, Soyeon, Allison J Ames, and Nicholas D Myers (2012). "A review of meta-analyses in education: Methodological strengths and weaknesses". *Review of Educational Research* 82.4, pp. 436–476 (cit. on p. 4).

Alamri, Hamdan, Victoria Lowell, William Watson, and Sunnie Lee Watson (2020). "Using Personalized Learning as an Instructional Approach to Motivate Learners in Online Higher Education: Learner Self-Determination and Intrinsic Motivation". *Journal of Research on Technology in Education* 52.3, pp. 322–352 (cit. on p. 37).

Albrecht, Christine, Ruben van de Vijver, and Christian Bellebaum (2023). "Learning new words via feedback—Association between feedback-locked ERPs and recall performance—An exploratory study". *Psychophysiology* 60.10, e14324 (cit. on p. 120).

Aleven, Vincent, Elizabeth A McLaughlin, R Amos Glenn, and Kenneth R Koedinger (2016). "Instruction based on adaptive learning technologies". *Handbook of research on learning and instruction* 2, pp. 522–560 (cit. on p. 6).

Aljabri, Sameer (2024). "Timing of feedback and retrieval practice: a laboratory study with EFL students". *Humanities and Social Sciences Communications* 11.1, pp. 1–10 (cit. on p. 120).

Aljawarneh, Shadi and Juan A. Lara (2021). "Data Science for Analyzing and Improving Educational Processes". *Journal of Computing in Higher Education* 33, pp. 545–550 (cit. on p. 37).

Allen, Justin P, Eui Kyung Kim, and Shane R Jimerson (2023). "Meta-Analyses and Systematic Reviews Advancing the Practice of School Psychology: The Imperative of Bringing Science to Practice". *School Psychology Review* 52.2, pp. 87–94 (cit. on p. 4).

Alshumaimeri, Yousif A (2023). "Understanding context: An essential factor for educational change success". *Contemporary Educational Research Journal* 13.1, pp. 11–19 (cit. on p. 9).

Anderson, John R, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier (1995). "Cognitive tutors: Lessons learned". *The journal of the learning sciences* 4.2, pp. 167–207 (cit. on p. 175).

Anderson, Lorin W and David R Krathwohl (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc. (cit. on p. 145).

Anderson, Richard C, Raymond W Kulhavy, and Thomas Andre (1972). "Conditions under which feedback facilitates learning from programmed lessons." *Journal of Educational Psychology* 63.3, p. 186 (cit. on p. 29).

Andler, Daniel (2008). "Sciences cognitives et éducation: une relation sérieuse". *Apprendre demain. Sciences cognitives et éducation à l'ère numérique, Paris: Hatier*, pp. 26–51 (cit. on p. 2).

Antal, Margit (2013). “On the Use of Elo Rating for Adaptive Assessment”. *Studia Universitatis Babes-Bolyai, Informatica* 58.1, pp. 29–41 (cit. on pp. 38, 43, 65).

Antoninis, Manos, Benjamin Alcott, Samaher Al Hadheri, Daniel April, Bilal Fouad Barakat, Marcela Barrios Rivera, et al. (2023). “Global Education Monitoring Report 2023: Technology in education: A tool on whose terms?” (Cit. on p. 5).

Anyichie, Aloysius C and Deborah L Butler (2023). “Examining culturally diverse learners’ motivation and engagement processes as situated in the context of a complex task”. *Frontiers in Education*. Vol. 8. Frontiers Media SA, p. 1041946 (cit. on p. 9).

Arnold, Kathleen M, Sharda Umanath, Kara Thio, Walter B Reilly, Mark A McDaniel, and Elizabeth J Marsh (2017). “Understanding the cognitive processes involved in writing to learn.” *Journal of Experimental Psychology: Applied* 23.2, p. 115 (cit. on p. 9).

Arroyo, Diana C and Yucel Yilmaz (2018). “An open for replication study: The role of feedback timing in synchronous computer-mediated communication”. *Language Learning* 68.4, pp. 942–972 (cit. on pp. 120, 142, 146).

Aslaksen, Karoline and Håvard Lorås (2018). “The modality-specific learning style hypothesis: a mini-review”. *Frontiers in psychology* 9, p. 1538 (cit. on p. 4).

Attali, Yigal (2014). “A Ranking Method for Evaluating Constructed Responses”. *Educational and Psychological Measurement* 74.5, pp. 795–808 (cit. on pp. 21, 37, 43, 65).

Attali, Yigal and Fabienne van der Kleij (2017). “Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving”. *Computers & Education* 110, pp. 154–169 (cit. on pp. 90, 120).

Ausubel, David Paul, Joseph Donald Novak, Helen Hanesian, et al. (1978). “Educational psychology: A cognitive view” (cit. on p. 27).

Author (2025). “Meta-analysis of the Impact of Feedback Timing on Learning Outcomes” (cit. on pp. 142, 143, 147).

Azevedo, Roger and Robert M Bernard (1995). “A meta-analysis of the effects of feedback in computer-based instruction”. *Journal of Educational Computing Research* 13.2, pp. 111–127 (cit. on pp. 30, 83, 84, 88, 116).

Badrinath, Anirudhan, Frederic Wang, and Zachary Pardos (2021). “pybkt: An Accessible Python Library of Bayesian Knowledge Tracing Models”. *arXiv preprint arXiv:2105.00385* (cit. on p. 19).

Bandura, Albert et al. (1986). “Social foundations of thought and action”. *Englewood Cliffs, NJ* 1986.23-28, p. 2 (cit. on p. 27).

Barnes, Jean M and Benton J Underwood (1959). “" Fate" of first-list associations in transfer theory.” *Journal of experimental psychology* 58.2, p. 97 (cit. on p. 31).

Barton, Jack, Kathrine Sofia Rallis, Amber Elyse Corrigan, Ella Hubbard, Antonia Round, Greta Portone, et al. (2021). “Medical students’ pattern of self-directed learning prior to and during the coronavirus disease 2019 pandemic period and its implications for Free Open Access Meducation within the United Kingdom”. *J Educ Eval Health Prof* 18.5 (cit. on p. 12).

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). “Fitting Linear Mixed-Effects Models Using lme4”. *Journal of Statistical Software* 67.1, pp. 1–48 (cit. on pp. 68, 155).

Bawa, Papia (2016). “Retention in online courses: Exploring issues and solutions—A literature review”. *SAGE Open* 6.1, pp. 1–11 (cit. on p. 11).

Belfer, Robert, Ekaterina Kochmar, and Iulian Vlad Serban (2022). “Raising student completion rates with adaptive curriculum and contextual bandits”. *International Conference on Artificial Intelligence in Education*. Springer, pp. 724–730 (cit. on p. 27).

Benvenuti, Martina, Angelo Cangelosi, Armin Weinberger, Elvis Mazzoni, Mariagrazia Benassi, Mattia Barbaresi, et al. (2023). “Artificial intelligence and human behavioral development: A perspective on new skills and competences acquisition for the educational context”. *Computers in Human Behavior* 148, p. 107903 (cit. on p. 75).

Bernacki, Matthew L, Meghan J Greene, and Nikki G Lobczowski (2021). “A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)?” *Educational Psychology Review* 33.4, pp. 1675–1715 (cit. on p. 6).

Bjork, Elizabeth L, Robert A Bjork, et al. (2011). “Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning”. *Psychology and the real world: Essays illustrating fundamental contributions to society* 2.59-68 (cit. on pp. 113, 174).

Bjork, Robert A (1994). "Memory and metamemory considerations in the training of human beings". *Metacognition: Knowing about knowing* 185.7.2, pp. 185–205 (cit. on p. 24).

Bjork, Robert A and Elizabeth L Bjork (2020). "Desirable difficulties in theory and practice." *Journal of Applied research in Memory and Cognition* 9.4, p. 475 (cit. on p. 24).

Bloom, Benjamin Samuel, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl (1964). *Taxonomy of educational objectives*. Vol. 2. Longmans, Green New York (cit. on pp. 87, 95, 113, 145).

Bock, Anna, Kristian Kniha, Evgeny Goloborodko, Martin Lemos, Anne Barbara Rittich, Stephan Christian Möhlhenrich, et al. (2021). "Effectiveness of face-to-face, blended and e-learning in teaching the application of local anaesthesia: a randomised study". *BMC medical education* 21, pp. 1–8 (cit. on p. 13).

Boulay, Ben du, Alexandra Poulovasillis, Wayne Holmes, and Manolis Mavrikis (2018). "Artificial Intelligence And Big Data Technologies To Close The Achievement Gap." (cit. on p. 10).

Boyle, James (2012). "Understanding the Nature of Experiments in Real-World Educational". *Handbook of implementation science for psychology in education*, p. 54 (cit. on p. 9).

Bradley, Robert H and Robert F Corwyn (2002). "Socioeconomic status and child development". *Annual review of psychology* 53.1, pp. 371–399 (cit. on p. 5).

Bressoux, Pascal, Laurent Lima, and Christian Monseur (2019). "Reducing the number of pupils in French first-grade classes: Is there evidence of contemporaneous and carryover effects?" *International Journal of Educational Research* 96, pp. 136–145 (cit. on p. 2).

Broeke, Nick ten, Abe Hofman, Joost Kruis, Susanne de Mooij, and Han van der Maas (2022). "Predicting and reducing quitting in online learning" (cit. on p. 27).

Brooks, Christopher and Craig Thompson (2017). "Predictive Modelling in Teaching and Learning". *Handbook of Learning Analytics*, pp. 61–68 (cit. on p. 37).

Brummer, Leonie, Hester de Boer, Jolien M Mouw, and Jan-Willem Strijbos (2024). "A meta-analysis of the effects of context, content, and task factors of digitally delivered instructional feedback on learning performance". *Learning Environments Research* 27.3, pp. 453–476 (cit. on pp. 29, 30, 85, 112).

Bryan, William Lowe and Noble Harter (1899). "Studies on the telegraphic language: The acquisition of a hierarchy of habits." *Psychological review* 6.4, p. 345 (cit. on p. 9).

Bryk, Anthony S, Louis M Gomez, Alicia Grunow, and Paul G LeMahieu (2015). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press (cit. on p. 4).

Buja, L Maximilian (2019). "Medical education today: all that glitters is not gold". *BMC medical education* 19, pp. 1–11 (cit. on p. 12).

Butler, Andrew C, Lisa K Fazio, and Elizabeth J Marsh (2011). "The hypercorrection effect persists over a week, but high-confidence errors return". *Psychonomic Bulletin & Review* 18, pp. 1238–1244 (cit. on pp. 31, 84, 143, 145).

Butler, Andrew C, Jeffrey D Karpicke, and Henry L Roediger III (2007). "The effect of type and timing of feedback on learning from multiple-choice tests." *Journal of Experimental Psychology: Applied* 13.4, p. 273 (cit. on pp. 84, 115, 120, 141, 142, 144–146).

Butler, Andrew C and Henry L Roediger (2008). "Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing". *Memory & cognition* 36.3, pp. 604–616 (cit. on pp. 29, 120, 141, 144).

Butler, Deborah L and Philip H Winne (1995). "Feedback and self-regulated learning: A theoretical synthesis". *Review of educational research* 65.3, pp. 245–281 (cit. on pp. 29, 83, 141).

Butterfield, Brady and Janet Metcalfe (2006). "The correction of errors committed with high confidence". *Metacognition and Learning* 1, pp. 69–84 (cit. on p. 28).

Cai, Zhihui, Yang Gui, Peipei Mao, Zhikeng Wang, Xin Hao, Xitao Fan, et al. (2023). "The effect of feedback on academic achievement in technology-rich learning environments (TREs): A meta-analytic review". *Educational Research Review* 39, p. 100521 (cit. on pp. 29, 30, 83, 85, 112).

Canals, Laia, Gisela Granena, Yucel Yilmaz, and Aleksandra Malicka (2021). "The relative effectiveness of immediate and delayed corrective feedback in video-based computer-mediated communication". *Language Teaching Research*, p. 13621688211052793 (cit. on pp. 112, 120).

Candel, Carmen, Ignacio Máñez, Raquel Cerdán, and Eduardo Vidal-Abarca (2021). "Delaying elaborated feedback within computer-based learning environments: The role of summative and question-based feedback". *Journal of Computer Assisted Learning* 37.4, pp. 1015–1029 (cit. on p. 120).

Candel, Carmen, Eduardo Vidal-Abarca, Raquel Cerdán, Marie Lippmann, and Susanne Narciss (2020). "Effects of timing of formative feedback in computer-assisted learning environments". *Journal of Computer Assisted Learning* 36.5, pp. 718–728 (cit. on p. 120).

Cao, Yang, Shao-Ying Gong, Zhen Wang, Yang Cheng, and Yan-Qing Wang (2022). "More challenging or more achievable? The impacts of difficulty and dominant goal orientation in leaderboards within educational gamification". *Journal of Computer Assisted Learning* 38.3, pp. 845–860 (cit. on p. 59).

Carpenter, Shana K (2014). "Spacing and interleaving of study and practice". *Applying the science of learning in education: Infusing psychological science into the curriculum*, pp. 131–141 (cit. on p. 113).

Carpenter, Shana K and Edward Vul (2011). "Delaying feedback by three seconds benefits retention of face–name pairs: The role of active anticipatory processing". *Memory & Cognition* 39, pp. 1211–1221 (cit. on pp. 115, 120, 142, 144, 146).

Cartwright, Nancy (2020). "What is meant by "rigour" in evidence-based educational policy and what's so good about it?" *The Evidential Basis of "Evidence-Based Education"*. Routledge, pp. 63–80 (cit. on p. 3).

Case, Susan M and David B Swanson (1998). *Constructing written test questions for the basic and clinical sciences*. National Board of Medical Examiners Philadelphia (cit. on p. 176).

Cepeda, Nicholas J, Noriko Coburn, Doug Rohrer, John T Wixted, Michael C Mozer, and Harold Pashler (2009). "Optimizing distributed practice: Theoretical analysis and practical implications". *Experimental psychology* 56.4, pp. 236–246 (cit. on pp. 84, 144).

Cepeda, Nicholas J., Edward Vul, Doug Rohrer, John T. Wixted, and Harold Pashler (2008). "Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention". *Psychological Science* 19.11, pp. 1095–1102 (cit. on pp. 54, 84, 115, 144, 145).

Chan, Wai-Lun and Dit-Yan Yeung (2021). "Clickstream knowledge tracing: Modeling how students answer interactive online questions". *LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 99–109 (cit. on pp. 37, 65).

Chanifah, Sabila, Rachmadita Andreswari, and Rokhman Fauzi (2021). "Analysis of student learning pattern in learning management system (LMS) using heuristic mining a process mining approach". *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*. IEEE, pp. 121–125 (cit. on p. 8).

Chen, Lijia, Pingping Chen, and Zhijian Lin (2020). "Artificial intelligence in education: A review". *Ieee Access* 8, pp. 75264–75278 (cit. on p. 175).

Chen, Xin, Huiyun Yuan, Yong Zhang, Dominique Bertrand, Gilbert Vicente, and Wenhui Zhang (2024). "Characteristics and considerations of French medical education". *Global Medical Education* 1.1, pp. 21–29 (cit. on p. 13).

Choffin, Benoıt, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie (2019). "DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills". *Proceedings of the Twelfth International Conference on Educational Data Mining (EDM 2019)*, pp. 29–38 (cit. on pp. 20, 21, 37, 53).

Chrysafiadi, Konstantina and Maria Virvou (2013). "Student modeling approaches: A literature review for the last decade". *Expert Systems with Applications* 40.11, pp. 4715–4729 (cit. on p. 19).

Cirigliano, Matthew M, Charles D Guthrie, and Martin V Pusic (2020). "Click-level learning analytics in an online medical education learning platform". *Teaching and learning in medicine* 32.4, pp. 410–421 (cit. on p. 13).

Clariana, Roy B, Daren Wagner, and Lucia C Roher Murphy (2000). "Applying a connectionist description of feedback timing". *Educational Technology Research and Development* 48.3, pp. 5–22 (cit. on pp. 84, 117, 121, 142, 144).

Clinton-Lisell, Virginia and Christine Litzinger (2024). "Is it really a neuromyth? A meta-analysis of the learning styles matching hypothesis". *Frontiers in Psychology* 15, p. 1428732 (cit. on p. 4).

Conole, Gráinne, Dragan Gašević, Phillip Long, and George Siemens (2011). "Message from the LAK 2011 general & program chairs". *International Learning Analytics & Knowledge Conference 2011*. Association for Computing Machinery (ACM) (cit. on pp. 7, 8).

Contrino, Monica F, Maribell Reyes-Millán, Patricia Vázquez-Villegas, and Jorge Membrillo-Hernández (2024). "Using an adaptive learning tool to improve student performance and satisfaction in online and face-to-face education for a more personalized approach". *Smart Learning Environments* 11.1, p. 6 (cit. on p. 6).

Cook, Thomas D (2002). "Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them". *Educational evaluation and policy analysis* 24.3, pp. 175–199 (cit. on p. 3).

Cook, Thomas D (2007). "Randomized experiments in education: Assessing the objections to doing them". *Economics of Innovation and New Technology* 16.5, pp. 331–355 (cit. on p. 3).

Corbett, Albert T. and John R. Anderson (1994). "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge". *User Modeling and User-Adapted Interaction* 4, pp. 253–278 (cit. on pp. 19, 37).

Corral, Daniel, Shana K Carpenter, and Sam Clingan-Siverly (2021). "The effects of immediate versus delayed feedback on complex concept learning". *Quarterly Journal of Experimental Psychology* 74.4, pp. 786–799 (cit. on pp. 86, 121).

Corrégé, Jean-Baptiste and Nicolas Michinov (2021). "Group size and peer learning: Peer discussions in different group size influence learning in a biology exercise performed on a tablet with stylus". *Frontiers in Education*. Vol. 6. Frontiers Media SA, p. 733663 (cit. on p. 9).

Csikszentmihalyi, Mihaly (1990). *Flow: The Psychology of Optimal Experience*. New York: Harper Row (cit. on p. 23).

Cutting, Joe, Sebastian Deterding, Simon Demediuk, and Nick Sephton (2023). "Difficulty-skill balance does not affect engagement and enjoyment: a pre-registered study using artificial intelligence-controlled difficulty". *Royal Society Open Science* 10.2, p. 220274 (cit. on p. 79).

Dawson, Shane, Dragan Gašević, George Siemens, and Srecko Joksimovic (2014). "Current state and future trends: A citation network analysis of the learning analytics field". *Proceedings of the fourth international conference on learning analytics and knowledge*, pp. 231–240 (cit. on p. 8).

De Ayala, Rafael Jaime (2013). *The Theory and Practice of Item Response Theory*. Guilford Publications (cit. on p. 20).

De Freitas, S.I. and J. Morgan (2015). "Will MOOCs transform learning and teaching in higher education? Engagement and course retention in online learning provision". *British Journal of Educational Technology* 46.3, pp. 455–471 (cit. on p. 11).

De Morais, Alana M., Joseana MFR Araujo, and Evandro B. Costa (2014). "Monitoring Student Performance Using Data Clustering and Predictive Modelling". *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. IEEE, pp. 1–8 (cit. on p. 57).

Deaton, Angus and Nancy Cartwright (2018). "Understanding and misunderstanding randomized controlled trials". *Social science & medicine* 210, pp. 2–21 (cit. on p. 3).

Dempsey, John V and Susan U Wager (1988). "A taxonomy for the timing of feedback in computer-based instruction". *Educational Technology* 28.10, pp. 20–25 (cit. on p. 86).

DEPP - Direction de l'Évaluation, de la Prospective et de la Performance (2022). *Évaluation de l'impact de la réduction de la taille des classes de CP et de CE1 en REP+ sur les résultats des élèves*. Tech. rep. Ministère de l'Éducation nationale (cit. on p. 4).

Deterding, Sebastian and Joe Cutting (2023). "Objective difficulty-skill balance impacts perceived balance but not behaviour: A test of flow and self-determination theory predictions". *Proceedings of the ACM on Human-Computer Interaction* 7.CHI PLAY, pp. 1179–1205 (cit. on p. 58).

Deunk, Marjolein I, Annemieke E Smale-Jacobse, Hester de Boer, Simone Doolaard, and Roel J Bosker (2018). "Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education". *Educational Research Review* 24, pp. 31–54 (cit. on p. 6).

Ding, Guozhu, Mailin Li, Shan Li, and Hao Wu (2024). "Exploring the impact of feedback timing on student performance in online testing". *Asia Pacific Education Review*, pp. 1–13 (cit. on p. 121).

Donkin, Rebecca, Elizabeth Askew, and Hollie Stevenson (2019). "Video feedback and e-Learning enhances laboratory skills and engagement in medical laboratory science students". *BMC medical education* 19, pp. 1–12 (cit. on p. 85).

Dore, Rebecca A and Jaclyn M Dynia (2020). "Technology and media use in preschool classrooms: Prevalence, purposes, and contexts". *Frontiers in Education*. Vol. 5. Frontiers Media SA, p. 600305 (cit. on p. 5).

Dost, Samiullah, Aleena Hossain, Mai Shehab, Aida Abdelwahed, and Lana Al-Nusair (2020). "Perceptions of medical students towards online teaching during the COVID-19 pandemic: a national cross-sectional survey of 2721 UK medical students". *BMJ open* 10.11, e042378 (cit. on p. 12).

Dumont, Hanna and Douglas D Ready (2023). "On the promise of personalized learning for educational equity". *Npj science of learning* 8.1, p. 26 (cit. on p. 5).

Dunbar, Norah E, Matthew L Jensen, Claude H Miller, Elena Bessarabova, Yu-Hao Lee, Scott N Wilson, et al. (2017). "Mitigation of cognitive bias with a serious game: Two experiments testing feedback timing and source". *International Journal of Game-Based Learning (IJGBL)* 7.4, pp. 86–100 (cit. on pp. 90, 122).

Ebbinghaus, Hermann (2013). "Memory: A Contribution to Experimental Psychology". *Annals of Neurosciences* 20.4, p. 155 (cit. on p. 53).

Eggen, Theo JHM, Fabienne M van der Kleij, and Caroline F Timmers (2011). "The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review". *Cadmo: giornale italiano di pedagogia sperimentale: 1, 2011*, pp. 21–38 (cit. on pp. 30, 142, 145, 146).

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder (1997). "Bias in meta-analysis detected by a simple, graphical test". *bmj* 315.7109, pp. 629–634 (cit. on p. 98).

Eglington, Luke G and Philip I Pavlik Jr (2023). "How to optimize student learning using student models that adapt rapidly to individual differences". *International Journal of Artificial Intelligence in Education* 33.3, pp. 497–518 (cit. on p. 19).

Elliot, A. J. and J. Harackiewicz (1994). "Goal setting, achievement orientation, and intrinsic motivation: A mediational analysis". *Journal of Personality and Social Psychology* 66, pp. 968–980 (cit. on p. 77).

Elo, Arpad E. and Sam Sloan (1978). *The Rating of Chessplayers: Past and Present* (cit. on pp. 21, 37, 60, 65).

Erdman, Matthew R and Jason CK Chan (2013). "Providing corrective feedback during retrieval practice does not increase retrieval-induced forgetting". *Journal of Cognitive Psychology* 25.6, pp. 692–703 (cit. on p. 122).

Erdmann, Julia and Nikol Rummel (2022). "The role of spontaneous recovery effects in the context of German orthography instruction methods with delayed correction." *Journal of Experimental Psychology: Applied* 28.1, p. 130 (cit. on p. 122).

Eyre, Heidi L (2007). "Keller's Personalized System of Instruction: Was it a Fleeting Fancy or is there a Revival on the Horizon?" *The Behavior Analyst Today* 8.3, p. 317 (cit. on p. 26).

Al-Fawakhiri, Naser, Sarosh Kayani, and Samuel D McDougle (2023). "Evidence of an optimal error rate for motor skill learning". *bioRxiv*, pp. 2023–07 (cit. on p. 26).

Filges, Trine, Christoffer Scavenius Sonne-Schmidt, and Bjørn Christian Viinholt Nielsen (2018). "Small class sizes for improving student achievement in primary and secondary schools: A systematic review". *Campbell Systematic Reviews* 14.1, pp. 1–107 (cit. on p. 2).

Fisher, Zachary and Elizabeth Tipton (2015). "robumeta: An R-package for robust variance estimation in meta-analysis". *arXiv preprint arXiv:1503.02220* (cit. on p. 98).

Fong, Carlton J, Erika A Patall, Ariana C Vasquez, and Sandra Stautberg (2019). "A meta-analysis of negative feedback on intrinsic motivation". *Educational Psychology Review* 31, pp. 121–162 (cit. on p. 29).

French Ministry for Higher Education and Research (2013). *Etudes Médicales* (cit. on p. 15).

Fu, Qing-Ke and Gwo-Jen Hwang (2018). "Trends in mobile technology-supported collaborative learning: A systematic review of journal publications from 2007 to 2016". *Computers & Education* 119, pp. 129–143 (cit. on pp. 29, 142).

Furey, William (2020). "THE STUBBORN MYTH OF" LEARNING STYLES"." *Education Next* 20.3 (cit. on p. 4).

Furlan, Raffaello, Mauro Gatti, Roberto Mene, Dana Shiffer, Chiara Marchiori, Alessandro Giaj Levra, et al. (2022). "Learning analytics applied to clinical diagnostic reasoning using a natural language processing–based virtual patient simulator: case study". *JMIR Medical Education* 8.1, e24372 (cit. on p. 13).

Fyfe, Emily R (2016). "Providing feedback on computer-based algebra homework in middle-school classrooms". *Computers in Human Behavior* 63, pp. 568–574 (cit. on pp. 9, 122).

Fyfe, Emily R and Bethany Rittle-Johnson (2016). "The benefits of computer-generated feedback for mathematics problem solving". *Journal of experimental child psychology* 147, pp. 140–151 (cit. on p. 122).

Gallego, F, R Molina, and F Llorens (2018). "Measuring the difficulty of activities for adaptive learning". *Universal Access in the Information Society. https://doi. org/10.1007/s10209-017-0552-x* (cit. on pp. 25, 26, 173).

García-Pérez, M. A. (1998). "Forced-choice Staircases with Fixed Step Sizes: Asymptotic and Small-sample Properties". *Vision Research* 38, pp. 1861–1881 (cit. on p. 58).

Gasevic, Dragan, Jelena Jovanovic, Abelardo Pardo, and Shane Dawson (2017). "Detecting learning strategies with analytics: Links with self-reported measures and academic performance". *Journal of Learning Analytics* 4.2, pp. 113–128 (cit. on p. 177).

Gašević, Dragan, Shane Dawson, and George Siemens (2015). "Let's not forget: Learning analytics are about learning". *TechTrends* 59, pp. 64–71 (cit. on p. 175).

Gersten, Russell M., Douglas W. Carnine, and Paul B. Williams (1982). "Measuring Implementation of a Structured Educational Model in an Urban School District: An Observational Approach". *Educational Evaluation and Policy Analysis* 4.1, pp. 67–79 (cit. on pp. 19, 57).

Ghosh, Aritra, Neil Heffernan, and Andrew S Lan (2020). "Context-aware attentive knowledge tracing". *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2330–2339 (cit. on p. 37).

Goda, Yoshiko (2004). *Feedback timing and learners' response confidence on learning English as a foreign language (EFL): Examining the effects of a computer-based feedback and assessment environment on EFL students' language acquisition*. Florida Institute of Technology (cit. on p. 122).

Goemaere, Sophie, Wim Beyers, Gert-Jan De Muynck, and Maarten Vansteenkiste (2018). "The paradoxical effect of long instructions on negative affect and performance: When, for whom and why do they backfire?" *Acta Astronautica* 147, pp. 421–430 (cit. on p. 22).

Golke, Stefanie, Tobias Dörfler, and Cordula Artelt (2015). "The impact of elaborated feedback on text comprehension within a computer-based assessment". *Learning and instruction* 39, pp. 123–136 (cit. on p. 141).

González-Brenes, José, Yun Huang, and Peter Brusilovsky (2014). "General Features in Knowledge Tracing to Model Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge". *The 7th International Conference on Educational Data Mining*. University of Pittsburgh, pp. 84–91 (cit. on p. 20).

Good, Thomas L. and Douglas A. Grouws (1977). "Teaching Effects: A Process-Product Study in Fourth-Grade Mathematics Classrooms". *Journal of Teacher Education* 28.3, pp. 49–54 (cit. on p. 57).

Goomas, David and Kurt Czupryn (2021). "Using a learning management system common template in teaching adult basic education: Opportunities and challenges". *Community College Journal of Research and Practice* 45.3, pp. 227–230 (cit. on p. 10).

Gopalan, Maithreyi, Kelly Rosinger, and Jee Bin Ahn (2020). "Use of quasi-experimental research designs in education research: Growth, promise, and challenges". *Review of Research in Education* 44.1, pp. 218–243 (cit. on p. 178).

Greenland, Sander, Mohammad Ali Mansournia, and Douglas G Altman (2016). "Sparse data bias: a problem hiding in plain sight". *bmj* 352 (cit. on p. 177).

Grimaldi, Phillip J and Jeffrey D Karpicke (2012). "When and why do retrieval attempts enhance subsequent encoding?" *Memory & cognition* 40, pp. 505–513 (cit. on pp. 143, 145).

Guzmán-Muñoz, Francisco J and Addie Johnson (2008). "Error feedback and the acquisition of geographical representations". *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 22.7, pp. 979–995 (cit. on pp. 122, 142).

Gweon, G-H, Hee-Sun Lee, Chad Dorsey, Robert Tinker, William Finzer, and Daniel Damelin (2015). "Tracking student progress in a game-like learning environment with a monte carlo bayesian knowledge tracing model". *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pp. 166–170 (cit. on pp. 37, 65).

Hall, G Stanley (1891). "The contents of children's minds on entering school". *The Pedagogical Seminary* 1.2, pp. 139–173 (cit. on p. 9).

Harrer, Mathias, Pim Cuijpers, Toshi Furukawa, and David Ebert (2021). *Doing meta-analysis with R: A hands-on guide*. Chapman and Hall/CRC (cit. on p. 116).

Hattie, John (1999). "Influences on student learning". *Inaugural lecture given on August* 2.1999, p. 21 (cit. on p. 141).

— (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge (cit. on pp. 57, 67).

Hattie, John and Shirley Clarke (2018). *Visible learning: feedback*. Routledge (cit. on p. 83).

Hattie, John and Helen Timperley (2007). "The power of feedback". *Review of educational research* 77.1, pp. 81–112 (cit. on pp. 29, 30, 83).

Hays, Matthew Jensen, Nate Kornell, and Robert A Bjork (2013). "When and why a failed test potentiates the effectiveness of subsequent study." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 39.1, p. 290 (cit. on p. 122).

Hedges, Larry V, Elizabeth Tipton, and Matthew C Johnson (2010). "Robust variance estimation in meta-regression with dependent effect size estimates". *Research synthesis methods* 1.1, pp. 39–65 (cit. on p. 98).

Heffernan, Neil T and Cristina Lindquist Heffernan (2014). "The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching". *International Journal of Artificial Intelligence in Education* 24, pp. 470–497 (cit. on p. 10).

Henderson, Carly (2021). "The effect of feedback timing on L2 Spanish vocabulary acquisition in synchronous computer-mediated communication". *Language Teaching Research* 25.2, pp. 185–208 (cit. on p. 122).

Henshaw, Florencia (2011). "Effects of feedback timing in SLA: A computer assisted study on the Spanish subjunctive". *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism*, pp. 85–99 (cit. on p. 122).

Horvath, Michael, Hailey A Herleman, and R Lee McKie (2006). "Goal orientation, task difficulty, and task interest: A multilevel analysis". *Motivation and emotion* 30, pp. 169–176 (cit. on p. 22).

Hu, Sumei (2024). "The effect of artificial intelligence-assisted personalized learning on student learning outcomes: A meta-analysis based on 31 empirical research papers". *Science Insights Education Frontiers* 24.1, pp. 3873–3894 (cit. on p. 7).

Huang, Zhenya, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, et al. (2017). "Question Difficulty Prediction for READING Problems in Standard Tests". *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1 (cit. on p. 26).

Hull, Clark L (1952). "A behavior system; an introduction to behavior theory concerning the individual organism." (cit. on pp. 83, 143).

Hunicke, Robin (2005). "The case for dynamic difficulty adjustment in games". *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*, pp. 429–433 (cit. on p. 27).

Institute of Education Sciences (2018). *Building evidence: Changes to the IES goal structure for FY 2019*. Blog post (cit. on p. 3).

Ismail, Shahrul Nizam, Suraya Hamid, Muneer Ahmad, Abdulellah Alaboudi, and Nz Jhanjhi (2021). "Exploring students engagement towards the learning management system (LMS) using learning analytics." *Computer systems science & engineering* 37.1 (cit. on p. 9).

Israel-Fishelson, Rotem and Arnon Hershkovitz (2021). "Micro-persistence and difficulty in a game-based learning environment for computational thinking acquisition". *Journal of Computer Assisted Learning* 37.3, pp. 839–850 (cit. on p. 79).

Iwaki, Nobuyoshi, Tomomi Nara, and Saeko Tanaka (2017). "Does delayed corrective feedback enhance acquisition of correct information?" *Acta Psychologica* 181, pp. 75–81 (cit. on pp. 31, 123, 145, 146).

Jacoby, Larry L, Christopher N Wahlheim, and Colleen M Kelley (2015). "Memory consequences of looking back to notice change: Retroactive and proactive facilitation." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41.5, p. 1282 (cit. on p. 31).

Jaehnig, Wendy and Matthew L Miller (2007). "Feedback types in programmed instruction: A systematic review". *The psychological record* 57.2, pp. 219–232 (cit. on p. 85).

Janelli, Maria and Anastasiya A Lipnevich (2021). "Effects of pre-tests and feedback on performance outcomes and persistence in Massive Open Online Courses". *Computers & Education* 161, p. 104076 (cit. on p. 85).

Jansen, Brenda RJ, Jolien Louwerse, Marthe Straatemeier, Sanne HG Van der Ven, Sharon Klinkenberg, and Han LJ Van der Maas (2013). "The influence of experiencing success in math on math anxiety, perceived math competence, and math performance". *Learning and Individual Differences* 24, pp. 190–197 (cit. on pp. 59, 75).

Johnson, Amy M, Jana Reisslein, and Martin Reisslein (2015). "Transitional feedback schedules during computer-based problem-solving practice". *Computers & Education* 81, pp. 270–280 (cit. on p. 123).

Kandemir, E. N., J.-J. Vie, A. Sanchez-Ayte, O. Palombi, and F. Ramus (2024). "Adaptation of the Multi-Concept Multivariate Elo Rating System to Medical Students' Training Data". *Proceedings of the Fourteenth International Conference on Learning Analytics and Knowledge (LAK 2024)* (cit. on pp. 65, 66).

Kapur, Manu (2014). "Productive failure in learning math". *Cognitive science* 38.5, pp. 1008–1022 (cit. on p. 28).

— (2016). "Examining productive failure, productive success, unproductive failure, and unproductive success in learning". *Educational Psychologist* 51.2, pp. 289–299 (cit. on p. 28).

Karaoglan Yilmaz, Fatma Gizem and Ramazan Yilmaz (2022). "Learning analytics intervention improves students' engagement in online learning". *Technology, Knowledge and Learning* 27.2, pp. 449–460 (cit. on p. 6).

Käser, Tanja, Severin Klingler, Alexander G. Schwing, and Markus Gross (2017). "Dynamic Bayesian Networks for Student Modeling". *IEEE Transactions on Learning Technologies* 10.4, pp. 450–462 (cit. on p. 19).

Kew, Si Na and Zaidatun Tasir (2022). "Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data". *Technology, Knowledge and Learning* 27.2, pp. 405–427 (cit. on p. 5).

Khajah, Mohammad, Robert V. Lindsey, and Michael C. Mozer (2016). "How Deep Is Knowledge Tracing?" *arXiv preprint arXiv:1604.02416* (cit. on p. 19).

Khiat, Henry and Silke Vogel (2022). "A self-regulated learning management system: Enhancing performance, motivation and reflection in learning". *Journal of University Teaching and Learning Practice* 19.2, pp. 43–59 (cit. on p. 10).

Kim, Jong Hae (2019). "Multicollinearity and misleading statistical results". *Korean journal of anesthesiology* 72.6, p. 558 (cit. on pp. 69, 163).

Klinkenberg, Sharon, Marthe Straatemeier, and Han LJ van der Maas (2011). "Computer Adaptive Practice of Maths Ability Using a New Item Response Model for On-the-Fly Ability and Difficulty Estimation". *Computers & Education* 57.2, pp. 1813–1824 (cit. on pp. 19, 27, 37, 43, 57, 65, 66).

Kluger, Avraham N and Angelo DeNisi (1996). "The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory." *Psychological bulletin* 119.2, p. 254 (cit. on p. 141).

Kmet, LM (2004). "Standard quality assessment criteria for evaluating primary research papers from a variety of fields". *Alberta Heritage Foundation for Medical Research Edmonton* (cit. on p. 93).

Knight, Justin B, B Hunter Ball, Gene A Brewer, Michael R DeWitt, and Richard L Marsh (2012). "Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention". *Journal of Memory and Language* 66.4, pp. 731–746 (cit. on p. 31).

Koedinger, Kenneth R, Julie L Booth, and David Klahr (2013). "Instructional complexity and the science to constrain it". *Science* 342.6161, pp. 935–937 (cit. on pp. 9, 173).

Koedinger, Kenneth R and Elizabeth A McLaughlin (2016). "Closing the Loop with Quantitative Cognitive Task Analysis." *International Educational Data Mining Society* (cit. on p. 9).

Koenka, Alison C, Lisa Linnenbrink-Garcia, Hannah Moshontz, Kayla M Atkinson, Carmen E Sanchez, and Harris Cooper (2021). "A meta-analysis on the impact of grades and comments on academic motivation and achievement: A case for written feedback". *Educational Psychology* 41.7, pp. 922–947 (cit. on p. 29).

Koh, Joyce Hwee Ling and Ben Kei Daniel (2022). "Shifting online during COVID-19: A systematic review of teaching and learning strategies and their outcomes". *International Journal of Educational Technology in Higher Education* 19.1, p. 56 (cit. on p. 5).

Kornell, Nate, Matthew Jensen Hays, and Robert A Bjork (2009). "Unsuccessful retrieval attempts enhance subsequent learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.4, p. 989 (cit. on pp. 28, 143, 145).

Kornell, Nate and Janet Metcalfe (2014). "The effects of memory retrieval, errors and feedback on learning." (cit. on p. 29).

Kostons, Danny, Tamara van Gog, and Fred Paas (2010). "Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners". *Computers & Education* 54.4, pp. 932–940 (cit. on pp. 22, 58).

Kourtali, Nektaria-Efstathia and Lais Borges (2023). "The effects of feedback timing on L2 development in written SCMC". *Computer Assisted Language Learning*, pp. 1–29 (cit. on p. 123).

Krutka, Daniel G, Christine Greenhow, Royce Kimmons, Luci Pangrazio, and Torrey Trust (2024). "The State of Educational Technology Research and Practice". *Handbook of Children and Screens: Digital Media, Development, and Well-Being from Birth Through Adolescence*. Springer Nature Switzerland Cham, pp. 523–528 (cit. on p. 5).

Kucirkova, Natalia, Libby Gerard, and Marcia C Linn (2021). "Designing personalised instruction: A research and design framework". *British Journal of Educational Technology* 52.5, pp. 1839–1861 (cit. on p. 10).

Kulhavy, Raymond W (1977). "Feedback in written instruction". *Review of educational research* 47.2, pp. 211–232 (cit. on p. 144).

Kulhavy, Raymond W and Richard C Anderson (1972). "Delay-retention effect with multiple-choice tests." *Journal of Educational Psychology* 63.5, p. 505 (cit. on pp. 84, 115, 142–145).

Kulik, James A and Chen-Lin C Kulik (1988). "Timing of feedback and verbal learning". *Review of educational research* 58.1, pp. 79–97 (cit. on pp. 30, 83, 84, 86–89, 113, 114, 141, 142, 144, 171).

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen (2020). *lmerTest: Tests in Linear Mixed Effects Models* (cit. on p. 68).

Laeeq, Kashif and Zulfiqar Ali Memon (2021). "Scavenge: An intelligent multi-agent based voice-enabled virtual assistant for LMS". *Interactive Learning Environments* 29.6, pp. 954–972 (cit. on p. 10).

Landis, J Richard and Gary G Koch (1977). "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". *Biometrics*, pp. 363–374 (cit. on p. 93).

Lavolette, Elizabeth, Charlene Polio, and Jimin Kahng (2015). "The accuracy of computer-assisted feedback and students' responses to it". *Language, Learning & Technology* 19.2 (cit. on pp. 90, 123).

Lavolette, Elizabeth HP (2014). *Effects of feedback timing and type on learning ESL grammar rules*. Michigan State University (cit. on p. 123).

Lee, Dabae, Yeol Huh, Chun-Yi Lin, and Charles Morgan Reigeluth (2022). "Personalized learning practice in US learner-centered schools". *Contemporary Educational Technology* 14.4, ep385 (cit. on p. 5).

Lee, Jung and Ok-Choon Park (2008). "Adaptive instructional systems". *Handbook of research on educational communications and technology*. Citeseer, pp. 469–484 (cit. on pp. 6, 37).

Lee, Lap-Kei, Simon KS Cheung, and Lam-For Kwok (2020). "Learning Analytics: Current Trends and Innovative Practices". *Journal of Computers in Education* 7, pp. 1–6 (cit. on p. 37).

Lefevre, David and Benita Cox (2017). "Delayed instructional feedback may be more effective, but is this contrary to learners' preferences?" *British Journal of Educational Technology* 48.6, pp. 1357–1367 (cit. on pp. 84, 117).

Leggett, Jack MI and Jennifer S Burt (2021). "Errors may not cue recall of corrective feedback: Evidence against the mediation hypothesis of the testing effect." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47.1, p. 65 (cit. on p. 31).

Leppink, Jimmie (2015). "Data analysis in medical education research: a multilevel perspective". *Perspectives on Medical Education* 4, pp. 14–24 (cit. on p. 164).

Li, Shaofeng, Yan Zhu, and Rod Ellis (2016). "The effects of the timing of corrective feedback on the acquisition of a new linguistic structure". *The Modern Language Journal* 100.1, pp. 276–295 (cit. on pp. 115, 142, 146).

Lim, Lisa-Angelique, Sheridan Gentili, Abelardo Pardo, Vitomir Kovanović, Alexander Whitelock-Wainwright, Dragan Gašević, et al. (2021). "What changes, and for whom? A study of the impact of learning analytics-based process feedback in a large course". *Learning and Instruction* 72, p. 101202 (cit. on p. 85).

Lin, Li-Chun, I-Chun Hung, Kinshuk, and Nian-Shing Chen (2019). "The impact of student engagement on learning outcomes in a cyber-flipped course". *Educational Technology Research and Development* 67, pp. 1573–1591 (cit. on p. 68).

Linden, Wim J. van der and Ronald K. Hambleton, eds. (2013). *Handbook of Modern Item Response Theory*. Springer Science & Business Media (cit. on pp. 20, 37, 65).

Lindsey, Robert V., Jeffery D. Shroyer, Harold Pashler, and Michael C. Mozer (2014). "Improving Students' Long-Term Knowledge Retention Through Personalized Review". *Psychological Science* 25.3, pp. 639–647 (cit. on pp. 20, 21, 37).

Lipsey, Mark W (2001). "Practical meta-analysis". *Thousand Oaks* (cit. on pp. 94, 97).

Loehr, Abbey M, Lisa K Fazio, and Bethany Rittle-Johnson (2020). "The role of recalling previous errors in middle-school children's learning". *British Journal of Educational Psychology* 90.4, pp. 997–1014 (cit. on p. 31).

Loibl, Katharina and Timo Leuders (2018). "Errors during exploration and consolidation-the effectiveness of productive failure as sequentially guided discovery learning". *Journal Fur Mathematik-Didaktik* 39.1, pp. 69–96 (cit. on p. 28).

— (2019). "How to make failure productive: Fostering learning from errors through elaboration prompts". *Learning and Instruction* 62, pp. 1–10 (cit. on p. 28).

Lomas, Derek, Kishan Patel, Jodi L Forlizzi, and Kenneth R Koedinger (2013). "Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 89–98 (cit. on pp. 59, 76, 78).

Lomas, J. D., K. Koedinger, N. Patel, S. Shodhan, N. Poonwala, and J. L. Forlizzi (2017). "Is difficulty overrated? The effects of choice, novelty and suspense on intrinsic motivation in educational games". *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1028–1039 (cit. on p. 79).

Loon, Mariëtte H van and Niamh S Oeri (2023). "Examining on-task regulation in school children: Interrelations between monitoring, regulation, and task performance." *Journal of educational psychology* 115.3, p. 446 (cit. on p. 176).

Louw, Byron C (2020). "Educating Physicians for the 21st Century: Learning from the Experiences of 'Systems Citizens'". PhD thesis. The Pennsylvania State University (cit. on p. 12).

Lu, Xiwen, Adam Sales, and Neil T Heffernan (2021). "Immediate Versus Delayed Feedback on Learning: Do People's Instincts Really Conflict With Reality?" *Journal of Higher Education Theory and Practice* 21.16 (cit. on pp. 123, 142).

Lu, Xiwen, Wei Wang, Benjamin A Motz, Weibing Ye, and Neil T Heffernan (2023). "Immediate text-based feedback timing on foreign language online assignments: How immediate should immediate feedback be?" *Computers and education open* 5, p. 100148 (cit. on p. 123).

Lüdecke, Daniel (2021). *sjPlot: Data Visualization for Statistics in Social Science* (cit. on p. 68).

Lüdecke, Daniel, Dominique Makowski, Mattan S. Ben-Shachar, Philip Waggoner, and Indrajeet Patil (2021). *performance: Assessment of Regression Models Performance* (cit. on p. 68).

Lujan, Heidi L and Stephen E DiCarlo (2006). "Too much teaching, not enough learning: what is the solution?" *Advances in physiology education* 30.1, pp. 17–22 (cit. on p. 12).

Ma, Wenting, Olusola O Adesope, John C Nesbit, and Qing Liu (2014). "Intelligent tutoring systems and learning outcomes: A meta-analysis." *Journal of educational psychology* 106.4, p. 901 (cit. on pp. 7, 37, 57, 175).

Maghsudi, Setareh, Andrew Lan, Jie Xu, and Mihaela van Der Schaar (2021). "Personalized education in the artificial intelligence era: what to expect next". *IEEE Signal Processing Magazine* 38.3, pp. 37–50 (cit. on p. 75).

Major, Louis, Gill A Francis, and Maria Tsapali (2021). "The effectiveness of technology-supported personalised learning in low-and middle-income countries: A meta-analysis". *British Journal of Educational Technology* 52.5, pp. 1935–1964 (cit. on pp. 6, 7, 78).

Martin, Florence, Yan Chen, Robert L Moore, and Carl D Westine (2020). "Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018". *Educational Technology Research and Development* 68, pp. 1903–1929 (cit. on p. 19).

Marzano, Robert J (2001). "A New Era of School Reform: Going Where the Research Takes Us." (cit. on p. 4).

Mason, B Jean and Roger Bruning (2001). "Providing feedback in computer-based instruction: What the research tells us". *Retrieved February* 15, p. 2007 (cit. on p. 87).

McClure, Larry, Susan Yonezawa, and Makeba Jones (2010). "Can School Structures Improve Teacher-Student Relationships? The Relationship between Advisory Programs, Personalization and Students' Academic Achievement." *Education Policy Analysis Archives* 18.17, n17 (cit. on p. 6).

Medical Colleges, American Association of (2022). *Year 2 Questionnaire 2022* (cit. on p. 12).

Mera, Yeray, Gabriel Rodríguez, and Eugenia Marin-Garcia (2022). "Unraveling the benefits of experiencing errors during learning: Definition, modulating factors, and explanatory theories". *Psychonomic bulletin & review* 29.3, pp. 753–765 (cit. on p. 28).

Mertens, Ute, Bridgid Finn, and Marlit Annalena Lindner (2022). "Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis." *Journal of Educational Psychology* 114.8, p. 1743 (cit. on pp. 30, 86, 87, 113).

Metcalfe, Janet (2017). "Learning from errors". *Annual review of psychology* 68.1, pp. 465–489 (cit. on p. 28).

Metcalfe, Janet and Teal S Eich (2019). "Memory and truth: correcting errors with true feedback versus overwriting correct answers with errors". *Cognitive research: principles and implications* 4, pp. 1–18 (cit. on p. 28).

Metcalfe, Janet, Nate Kornell, and Bridgid Finn (2009). "Delayed versus immediate feedback in children's and adults' vocabulary learning". *Memory & cognition* 37.8, pp. 1077–1087 (cit. on pp. 84, 90, 115, 123, 141, 142, 144, 145).

Metcalfe, Janet and David B Miele (2014). "Hypercorrection of high confidence errors: Prior testing both enhances delayed performance and blocks the return of the errors". *Journal of Applied Research in Memory and Cognition* 3.3, pp. 189–197 (cit. on p. 31).

Metcalfe, Janet and Judy Xu (2018). "Learning from one's own errors and those of others". *Psychonomic Bulletin & Review* 25, pp. 402–408 (cit. on pp. 84, 143, 145).

Metcalfe, Janet, Judy Xu, Matti Vuorre, Robert Siegler, Dylan Wiliam, and Robert A Bjork (2025). "Learning from errors versus explicit instruction in preparation for a test that counts". *British Journal of Educational Psychology* 95.1, pp. 11–25 (cit. on p. 28).

Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (2021a). *Arrêté du 21 décembre 2021 relatif à l'organisation des épreuves nationales donnant accès au troisième cycle des études de médecine*. Online (cit. on p. 14).

— (2021b). *Décret n° 2021-1156 du 7 septembre 2021 relatif à l'accès au troisième cycle des études de médecine*. Online (cit. on p. 14).

Mirari, Kaitlynn (2022). "The effectiveness of adaptive learning systems in personalized education". *Journal of Education Review Provision* 2.3, pp. 107–115 (cit. on p. 6).

Mohamed, Saad and F Adnan (2020). "Feedback in Computer-Assisted Language Learning: A Meta-Analysis." *Tesl-Ej* 24.2, n2 (cit. on p. 83).

Mohsen, Mohammed Ali (2022). "Computer-mediated corrective feedback to improve L2 writing skills: A meta-analysis". *Journal of Educational Computing Research* 60.5, pp. 1253–1276 (cit. on p. 83).

Mondigo, Lynnard and Demelo Madrazo Lao (2017). "E-learning for introductory Computer Science concept on recursion applying two types of feedback methods in the learning assessment". *Asian Association of Open Universities Journal* 12.2, pp. 218–229 (cit. on p. 142).

Morrison, Gary R, Steven M Ross, Mala Gopalakrishnan, and Jason Casey (1995). "The effects of feedback and incentives on achievement in computer-based instruction". *Contemporary Educational Psychology* 20.1, pp. 32–50 (cit. on p. 123).

Morrison, Keith (2019). "Realizing the promises of replication studies in education". *Educational Research and Evaluation* 25.7-8, pp. 412–441 (cit. on p. 9).

— (2020). *Taming randomized controlled trials in education: Exploring key claims, issues and debates*. Routledge (cit. on p. 3).

Mory, Edna Holland (2013). "Feedback research revisited". *Handbook of research on educational communications and technology*. Routledge, pp. 738–776 (cit. on pp. 29, 86, 112, 141, 142, 144, 171).

Motz, Benjamin A, Paulo F Carvalho, Joshua R de Leeuw, and Robert L Goldstone (2018). "Embedding experiments: Staking causal inference in authentic educational contexts". *Journal of Learning Analytics* 5.2, pp. 47–59 (cit. on p. 9).

Mozer, Michael C. and Robert V. Lindsey (2016). "Predicting and Improving Memory Retention". *Big Data in Cognitive Science* 34 (cit. on p. 20).

Mullaney, Kellie M, Shana K Carpenter, Courtney Grotenhuis, and Steven Burianek (2014). "Waiting for feedback helps if you want to know the answer: The role of curiosity in the delay-of-feedback benefit". *Memory & Cognition* 42, pp. 1273–1284 (cit. on p. 123).

Mullet, Hillary G, Andrew C Butler, Berenice Verdin, Ricardo von Borries, and Elizabeth J Marsh (2014). "Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately". *Journal of Applied Research in Memory and Cognition* 3.3, pp. 222–229 (cit. on pp. 84, 117, 124, 141).

Nakata, Tatsuya (2015). "Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning?" *Language Teaching Research* 19.4, pp. 416–434 (cit. on pp. 90, 112, 115, 124, 145).

Narciss, Susanne (2008). "Feedback strategies for interactive learning tasks". *Handbook of research on educational communications and technology*. Routledge, pp. 125–143 (cit. on p. 30).

— (2013). "Designing and evaluating tutoring feedback strategies for digital learning". *Digital Education Review* 23, pp. 7–26 (cit. on pp. 29, 142).

— (2017). "Conditions and effects of feedback viewed through the lens of the interactive tutoring feedback model". *Scaling up assessment for learning in higher education*, pp. 173–189 (cit. on p. 30).

Nathan, Mitchell J, Ana C Stephens, DK Masarik, Martha W Alibali, and Kenneth R Koedinger (2002). "Representational fluency in middle school: A classroom study". *Proceedings of the twenty-fourth annual meeting of the North American chapter of the International Group for the Psychology of Mathematics Education*. Vol. 1. ERIC Clearinghouse for Science, Mathematics and Environmental Education . . ., pp. 462–472 (cit. on pp. 113, 117, 142).

Naumann, Johannes (2019). "The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment". *Frontiers in Psychology* 10, p. 1429 (cit. on p. 176).

Newton, Philip M and Atharva Salvi (2020). "How common is belief in the learning styles neuromyth, and does it matter? A pragmatic systematic review". *Frontiers in Education*. Vol. 5. Frontiers, p. 602451 (cit. on p. 4).

Ninaus, Manuel, Korbinian Moeller, Jake McMullen, and Kristian Kiili (2017). "Acceptance of game-based learning and intrinsic motivation as predictors for learning success and flow experience". *International Journal of Serious Games* 4.3, pp. 15–30 (cit. on p. 173).

Nkhoma, Mathews, Narumon Sriratanaviriyakul, Hiep Pham Cong, and Tri Khai Lam (2014). "Examining the mediating role of learning engagement, learning process and learning experience on the learning outcomes through localized real case studies". *Education+ Training* 56.4, pp. 287–302 (cit. on p. 177).

Nunn, Kristen, Robert Creighton, Victoria Tilton-Bolowsky, Yael Arbel, and Sofia Vallila-Rohter (2024). "The effect of feedback timing on category learning and feedback processing in younger and older adults". *Frontiers in Aging Neuroscience* 16, p. 1404128 (cit. on p. 124).

O'neill, Marianne, Richard A Rasor, and Wayne R Bartz (1976). "Immediate retention of objective test answers as a function of feedback complexity". *The Journal of Educational Research* 70.2, pp. 72–75 (cit. on p. 145).

Opitz, Bertram, Nicola K Ferdinand, and Axel Mecklinger (2011). "Timing matters: the impact of immediate and delayed feedback on artificial language learning". *Frontiers in human neuroscience* 5, p. 8 (cit. on p. 124).

Page, Matthew J, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, et al. (2021). "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". *bmj* 372 (cit. on p. 88).

Palombi, Olivier, Fabrice Jouanot, Nafissetou Nziengam, Behrooz Omidvar-Tehrani, Marie-Christine Rousset, and Adam Sanchez (2019). "OntoSIDES: Ontology-based student progress monitoring on the national evaluation system of French Medical Schools". *Artificial intelligence in medicine* 96, pp. 59–67 (cit. on p. 15).

Papoušek, Jan and Radek Pelánek (2015). "Impact of Adaptive Educational System Behaviour on Student Motivation". *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings*. Vol. 17. Madrid, Spain: Springer International Publishing (cit. on p. 37).

— (2017). "Should We Give Learners Control Over Item Difficulty?" *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pp. 299–303 (cit. on pp. 37, 43, 58).

Papoušek, Jan, Radek Pelánek, and Vít Stanislav (2014). "Adaptive Practice of Facts in Domains with Varied Prior Knowledge". *Educational Data Mining 2014* (cit. on pp. 38, 43, 65).

Papoušek, Jan, Vít Stanislav, and Radek Pelánek (2016a). "Evaluation of an Adaptive Practice System for Learning Geography Facts". *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pp. 134–142 (cit. on p. 57).

— (2016b). "Impact of Question Difficulty on Engagement and Learning". *Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings*. Vol. 13. Springer International Publishing, pp. 267–272 (cit. on pp. 22, 26, 59, 77–79, 177).

Park, Ok-choon and Jung Lee (2003). "Adaptive Instructional Systems". *Educational Technology Research and Development* 25, pp. 651–684 (cit. on p. 57).

Pashler, Harold, Patrice M Bain, Brian A Bottge, Arthur Graesser, Kenneth Koedinger, Mark McDaniel, et al. (2007). "Organizing Instruction and Study to Improve Student Learning. IES Practice Guide. NCER 2007-2004." *National Center for Education Research* (cit. on p. 8).

Pashler, Harold, Nicholas J Cepeda, John T Wixted, and Doug Rohrer (2005). "When does feedback facilitate learning of words?" *Journal of experimental psychology: Learning, Memory, and Cognition* 31.1, p. 3 (cit. on p. 141).

Patall, Erika A, Sophia Hooper, Ariana C Vasquez, Keenan A Pituch, and Rebecca R Steingut (2018). "Science class is too hard: Perceived difficulty, disengagement, and the role of teacher autonomy support from a daily diary perspective". *Learning and Instruction* 58, pp. 220–231 (cit. on p. 22).

Pavlik Jr, Philip I., Hao Cen, and Kenneth R. Koedinger (2009). *Performance Factors Analysis–A New Alternative to Knowledge Tracing*. URL: https://www.learnlab.org/research/wiki/images/2/2c/PavlikCenKoedinger-ITS2009.pdf (cit. on p. 20).

Pei, Leisi and Hongbin Wu (2019). "Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis". *Medical education online* 24.1, p. 1666538 (cit. on p. 13).

Pelánek, Radek (2014). "Application of Time Decay Functions and the Elo System in Student Modeling". *Educational Data Mining 2014* (cit. on pp. 38, 43).

— (2016). “Applications of the Elo Rating System in Adaptive Educational Systems”. *Computers & Education* 98, pp. 169–179 (cit. on pp. 37, 38, 43, 44, 65, 66).

— (2017). “Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques”. *User modeling and user-adapted interaction* 27, pp. 313–350 (cit. on p. 19).

Pelánek, Radek, Jan Papoušek, Jakub Řihák, Vít Stanislav, and Jiří Nižnan (2017). “Elo-based learner modeling for the adaptive practice of facts”. *User Modeling and User-Adapted Interaction* 27.1, pp. 89–118 (cit. on pp. 19, 21, 27, 57).

Peng, Hongchao, Shanshan Ma, and Jonathan Michael Spector (2019). “Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment”. *Smart Learning Environments* 6.1, pp. 1–14 (cit. on p. 6).

Perttula, Arttu, Kristian Kiili, Antero Lindstedt, and Pauliina Tuomi (2017). “Flow experience in game based learning–a systematic literature review”. *International Journal of Serious Games* 4.1, pp. 57–72 (cit. on p. 173).

Phye, Gary D and Thomas Andre (1989). “Delayed retention effect: Attention, perseveration, or both?” *Contemporary Educational Psychology* 14.2, pp. 173–185 (cit. on pp. 84, 144).

Piaget, Jean (1952). “Introduction: The biological problem of intelligence.” (cit. on p. 28).

Piech, Chris, Joel Bassen, Jennifer Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, et al. (2015). “Deep Knowledge Tracing”. *Advances in Neural Information Processing Systems (NIPS)*, pp. 505–513 (cit. on pp. 37, 65).

Plooy, Eileen du, Daleen Casteleijn, and Denise Franzsen (2024). “Personalized adaptive learning in higher education: a scoping review of key characteristics and impact on academic performance and engagement”. *Heliyon* (cit. on pp. 6, 177).

Popovic, Natasa, Tomo Popovic, Isidora Rovcanin Dragovic, and Oleg Cmiljanic (2018). “A Moodle-based blended learning solution for physiology education in Montenegro: a case study”. *Advances in physiology education* 42.1, pp. 111–117 (cit. on p. 13).

Potts, Rosalind and David R Shanks (2014). “The benefit of generating errors during learning.” *Journal of Experimental Psychology: General* 143.2, p. 644 (cit. on p. 28).

Power, Jason (2019). “The influence of task difficulty on engagement, performance and self-efficacy”. *Explorations in Technology Education Research: Helping Teachers Develop Research Informed Practice*, pp. 157–169 (cit. on pp. 22, 57).

Pyc, Mary A and Katherine A Rawson (2010). “Why testing improves memory: Mediator effectiveness hypothesis”. *Science* 330.6002, pp. 335–335 (cit. on p. 31).

Quinn, Paul (2014). “Delayed versus immediate corrective feedback on orally produced passive errors in English”. PhD thesis. University of Toronto Toronto (cit. on p. 112).

Rahmani, A.M., W. Groot, and H. Rahmani (2024). “Dropout in online higher education: a systematic literature review”. *International Journal of Educational Technology in Higher Education* 21.1 (cit. on p. 11).

Rasch, Georg (1960). *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests* (cit. on pp. 20, 37, 65).

Redding, Sam (2013). “Getting Personal: The Promise of Personalized Learning.” *Center on Innovations in Learning, Temple University* (cit. on p. 6).

Roediger III, Henry L and Mary A Pyc (2012). “Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice”. *Journal of Applied Research in Memory and Cognition* 1.4, pp. 242–248 (cit. on p. 8).

Roediger III, Henry L. and Jeffrey D. Karpicke (2006). “The Power of Testing Memory: Basic Research and Implications for Educational Practice”. *Perspectives on Psychological Science* 1.3, pp. 181–210 (cit. on p. 57).

Rohrer, Doug, Kelli Taylor, Harold Pashler, John T Wixted, and Nicholas J Cepeda (2005). “The effect of overlearning on long-term retention”. *Applied Cognitive Psychology* 19.3, pp. 361–374 (cit. on p. 115).

Romero, Cristóbal, Sebastián Ventura, Eva L Gibaja, Cesar Hervás, and Francisco Romero (2006). “Web-based Adaptive Training Simulator System for Cardiac Life Support”. *Artificial Intelligence in Medicine* 38.1, pp. 67–78 (cit. on p. 58).

Ronimus, Miia, Janne Kujala, Asko Tolvanen, and Heikki Lyytinen (2014). “Children’s engagement during digital game-based learning of reading: The effects of time, rewards, and challenge”. *Computers & Education* 71, pp. 237–246 (cit. on p. 78).

Roper, WJ (1977). "Feedback in computer assisted instruction". *Programmed learning and educational technology* 14.1, pp. 43–49 (cit. on p. 86).

Rosenshine, Barak (2012). "Principles of Instruction: Research-Based Strategies That All Teachers Should Know". *American Educator* 36.1, p. 12 (cit. on pp. 26, 57, 75).

Ruan, Sherry, Wei Wei, and James Landay (2021). "Variational deep knowledge tracing for language learning". *LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 323–332 (cit. on p. 37).

Rubin, Donald B (1992). "Meta-analysis: literature synthesis or effect-size surface estimation?" *Journal of Educational Statistics* 17.4, pp. 363–374 (cit. on p. 98).

Ryan, Anna T, Terry Judd, Carey Wilson, Douglas P Larsen, Simone Elliott, Kulamakan Kulasegaram, et al. (2024). "Timing's not everything: Immediate and delayed feedback are equally beneficial for performance in formative multiple-choice testing". *Medical education* 58.7, pp. 838–847 (cit. on p. 124).

Sadler, D Royce (1989). "Formative assessment and the design of instructional systems". *Instructional science* 18.2, pp. 119–144 (cit. on p. 83).

Salden, Ron JCM, Fred Paas, and Jeroen JG Van Merriënboer (2006). "Personalised Adaptive Task Selection in Air Traffic Control: Effects on". *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48.1, pp. 120–130 (cit. on p. 58).

Saltzman, Irving J (1951). "Delay of reward and human verbal learning." *Journal of experimental psychology* 41.6, p. 437 (cit. on pp. 83, 143).

Sampayo-Vargas, Sandra, Chris J. Cope, Zhen He, and Graeme J. Byrne (2013). "The Effectiveness of Adaptive Difficulty Adjustments on Students' Motivation and Learning in an Educational Computer Game". *Computers & Education* 69, pp. 452–462 (cit. on pp. 22, 57, 58).

Samsonau, Sergey V (2018). "Computer simulations combined with experiments for a calculus-based physics laboratory course". *Physics Education* 53.5, p. 055013 (cit. on p. 9).

Schneider, Michael and Franzis Preckel (2017). "Variables associated with achievement in higher education: A systematic review of meta-analyses." *Psychological bulletin* 143.6, p. 565 (cit. on p. 173).

Schroth, Marvin L (1992). "The effects of delay of feedback on a delayed concept formation transfer task". *Contemporary educational psychology* 17.1, pp. 78–82 (cit. on p. 93).

— (1995). "Variable delay of feedback procedures and subsequent concept formation transfer". *The Journal of General Psychology* 122.4, pp. 393–399 (cit. on p. 93).

Schroth, Marvin L and Elissa Lund (1993). "Role of delay of feedback on subsequent pattern recognition transfer tasks". *Contemporary Educational Psychology* 18.1, pp. 15–22 (cit. on p. 93).

Schütt, Andreas, Tobias Huber, Jasmin Nasir, Cristina Conati, and Elisabeth André (2023). "Does Difficulty Even Matter? Investigating Difficulty Adjustment and Practice Behavior in an Open-Ended Learning Task". *arXiv preprint arXiv:2311.01934* (cit. on pp. 57, 58).

Schuwirth, Lambert WT and Cees PM Van der Vleuten (2011). "Programmatic assessment: from assessment of learning to assessment for learning". *Medical teacher* 33.6, pp. 478–485 (cit. on p. 176).

Segal, Avi, Yossi Ben David, Joseph Jay Williams, Kobi Gal, and Yaar Shalom (2018). "Combining difficulty ranking with multi-armed bandits to sequence educational content". *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part II 19*. Springer, pp. 317–321 (cit. on p. 27).

Segouin, C., J. Jouquan, and B. Hodges (2007). "Country report: medical education in France". *Medical Education* 41.3, pp. 295–301 (cit. on p. 13).

Seufert, Tina (2020). "Building bridges between self-regulation and cognitive load—an invitation for a broad and differentiated attempt". *Educational Psychology Review* 32.4, pp. 1151–1162 (cit. on p. 176).

Sheather, Simon (2009). *A modern approach to regression with R*. Springer Science & Business Media (cit. on pp. 69, 163).

Shin, Dongmin, Yugeun Shim, Hangyeol Yu, Seewoo Lee, Byungsoo Kim, and Youngduck Choi (2021). "Saint+: Integrating temporal features for ednet correctness prediction". *LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 490–496 (cit. on p. 37).

Shintani, Natsuko and Scott Aubrey (2016). "The effectiveness of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment". *The Modern Language Journal* 100.1, pp. 296–319 (cit. on p. 124).

Shirah, Julie F and Pooja G Sidney (2023). "Computer-based feedback matters when relevant prior knowledge is not activated". *Learning and Instruction* 87, p. 101796 (cit. on p. 125).

Shute, Valerie J (2008). "Focus on formative feedback". *Review of educational research* 78.1, pp. 153–189 (cit. on pp. 29, 30, 83, 85–87, 95, 113, 114, 141, 142).

Simpson, Amber, Adam V Maltese, Alice Anderson, and Euisuk Sung (2020). "Failures, errors, and mistakes: A systematic review of the literature". *Mistakes, errors and failures across cultures: Navigating potentials*, pp. 347–362 (cit. on p. 27).

Simpson, Ormond (2013). "Student retention in distance education: are we failing our students?" *Open Learning: The Journal of Open, Distance and e-Learning* 28.2, pp. 105–119 (cit. on p. 11).

Sinha, Neha and Arnold Lewis Glass (2015). "Delayed, but not immediate, feedback after multiple-choice questions increases performance on a subsequent short-answer, but not multiple-choice, exam: Evidence for the dual-process theory of memory". *The Journal of general psychology* 142.2, pp. 118–134 (cit. on p. 125).

Sinha, Tanmay, Patrick Jermann, Nan Li, and Pierre Dillenbourg (2014). "Your Click Decides Your Fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions". *arXiv preprint arXiv:1407.7131* (cit. on p. 57).

Sitzman, Danielle M, Matthew G Rhodes, and Sarah K Tauber (2014). "Prior knowledge is more predictive of error correction than subjective confidence". *Memory & Cognition* 42, pp. 84–96 (cit. on p. 125).

Skinner, Burrhus Frederic (1958). "Teaching Machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction." *Science* 128.3330, pp. 969–977 (cit. on p. 142).

— (1965). *Science and human behavior*. 92904. Simon and Schuster (cit. on pp. 27, 83, 143).

Slavin, Robert E (2002). "Evidence-based education policies: Transforming educational practice and research". *Educational researcher* 31.7, pp. 15–21 (cit. on p. 9).

— (2020). "How evidence-based reform will transform research and practice in education". *Educational Psychologist* 55.1, pp. 21–31 (cit. on p. 2).

Slavin, Robert E and Alan CK Cheung (2017). "Lessons learned from large-scale randomized experiments". *Journal of Education for Students Placed at Risk (JESPAR)* 22.4, pp. 253–259 (cit. on pp. 2, 3).

Smith, Troy A and Daniel R Kimball (2010). "Learning from feedback: Spacing and the delay–retention effect." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36.1, p. 80 (cit. on pp. 125, 142, 144).

Smits, Marieke HSB, Jo Boon, Dominique MA Sluijsmans, and Tamara Van Gog (2008). "Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge". *Interactive Learning Environments* 16.2, pp. 183–193 (cit. on pp. 86, 125).

Strong, Brian and Frank Boers (2019). "Weighing up exercises on phrasal verbs: Retrieval versus trial-and-error practices". *The Modern Language Journal* 103.3, pp. 562–579 (cit. on p. 125).

Surber, John R and Richard C Anderson (1975). "Delay-retention effect in natural classroom settings." *Journal of Educational Psychology* 67.2, p. 170 (cit. on p. 144).

Swann, Christian, Richard Keegan, Lee Crust, and David Piggott (2016). "Psychological states underlying excellent performance in professional golfers:"Letting it happen" vs."making it happen"". *Psychology of Sport and Exercise* 23, pp. 101–113 (cit. on p. 24).

Swart, Elise K, Thijs MJ Nielen, and Maria T Sikkema-de Jong (2019). "Supporting learning from text: A meta-analysis on the timing and content of effective feedback". *Educational Research Review* 28, p. 100296 (cit. on pp. 30, 85–88, 113, 116, 141).

Sweller, John (1988). "Cognitive load during problem solving: Effects on learning". *Cognitive science* 12.2, pp. 257–285 (cit. on pp. 24, 114).

— (2011). "Cognitive load theory". *Psychology of learning and motivation*. Vol. 55. Elsevier, pp. 37–76 (cit. on p. 83).

Sweller, John, Jeroen JG Van Merriënboer, and Fred Paas (2019). "Cognitive architecture and instructional design: 20 years later". *Educational psychology review* 31, pp. 261–292 (cit. on p. 114).

Swindell, Linda K and Walter F Walls (1993). "Response confidence and the delay retention effect". *Contemporary Educational Psychology* 18.3, pp. 363–375 (cit. on pp. 144, 145).

Tabibian, Behzad, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez (2019). "Enhancing human learning via spaced repetition optimization". *Proceedings of the National Academy of Sciences* 116.10, pp. 3988–3993 (cit. on pp. 6, 37).

Tanaka, Pedro, Yoon Soo Park, Jay Roby, Kyle Ahn, Clinton Kakazu, Ankeet Udani, et al. (2021). "Milestone learning trajectories of residents at five anesthesiology residency programs". *Teaching and Learning in Medicine* 33.3, pp. 304–313 (cit. on p. 13).

Tanaka, Saeko, Makoto Miyatani, and Nobuyoshi Iwaki (2019). "Response format, not semantic activation, influences the failed retrieval effect". *Frontiers in Psychology* 10, p. 599 (cit. on p. 125).

Tanner-Smith, Emily E and Elizabeth Tipton (2014). "Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS". *Research synthesis methods* 5.1, pp. 13–30 (cit. on p. 98).

Taxipulati, Sayipujiamali and Hai-Dong Lu (2021). "The influence of feedback content and feedback time on multimedia learning achievement of college students and its mechanism". *Frontiers in Psychology* 12, p. 706821 (cit. on pp. 86, 112, 126).

Terrace, Herbert S (1963). "Discrimination learning with and without "errors" 1". *Journal of the experimental analysis of behavior* 6.1, pp. 1–27 (cit. on p. 27).

Timmers, Caroline and Bernard Veldkamp (2011). "Attention paid to feedback provided by a computer-based assessment for learning on information literacy". *Computers & Education* 56.3, pp. 923–930 (cit. on p. 117).

Timmers, Caroline F, Amber Walraven, and Bernard P Veldkamp (2015). "The effect of regulation feedback in a computer-based formative assessment on information problem solving". *Computers & education* 87, pp. 1–9 (cit. on pp. 29, 30).

Turnbull, D, R Chugh, and J Luck (2019). *Learning management systems: An overview In Tatnall A.(Ed.), Encyclopedia of Education and Information Technologie* (cit. on p. 6).

Van der Kleij, Fabienne M, Theo JHM Eggen, Caroline F Timmers, and Bernard P Veldkamp (2012). "Effects of feedback in a computer-based assessment for learning". *Computers & Education* 58.1, pp. 263–272 (cit. on pp. 84, 85, 90, 117, 126).

Van der Kleij, Fabienne M, Remco CW Feskens, and Theo JHM Eggen (2015). "Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis". *Review of educational research* 85.4, pp. 475–511 (cit. on pp. 30, 83, 85–88, 97, 113, 141, 142, 146).

Van Ginkel, Stan, Judith Gulikers, Harm Biemans, Omid Noroozi, Mila Roozen, Tom Bos, et al. (2019). "Fostering oral presentation competence through a virtual reality-based task for delivering feedback". *Computers & Education* 134, pp. 78–97 (cit. on pp. 29, 142).

VanLehn, Kurt (2011). "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems". *Educational psychologist* 46.4, pp. 197–221 (cit. on pp. 7, 37).

Vermeiren, Hanke, Joost Kruis, Maria Bolsinova, Han LJ van der Maas, and Abe D Hofman (2025). "Psychometrics of an Elo-based Large-Scale Online Learning System". *Computers and Education: Artificial Intelligence*, p. 100376 (cit. on pp. 21, 22, 176).

Vie, Jill-Jênn and Hisashi Kashima (2019). "Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing". *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 750–757 (cit. on p. 46).

Voice, Alison and Arran Stirton (2020). "Spaced Repetition: Towards More Effective Learning in STEM." *New Directions in the Teaching of Physical Sciences* 15.1, n1 (cit. on p. 6).

Vygotsky, Lev (1978). "Interaction Between Learning and Development". *Readings on the Development of Children* 23.3, pp. 34–41 (cit. on pp. 57, 75).

Vygotsky, Lev Semenovich and Michael Cole (1978). *Mind in society: Development of higher psychological processes*. Harvard university press (cit. on pp. 23, 174).

Wang, Shuai, Claire Christensen, Wei Cui, Richard Tong, Louise Yarnall, Linda Shear, et al. (2023). "When adaptive learning is effective learning: comparison of an adaptive learning system to teacher-led instruction". *Interactive Learning Environments* 31.2, pp. 793–803 (cit. on p. 27).

Wang, W., L. Guo, L. He, and Y.J. Wu (2019). "Effects of social-interactive engagement on the dropout ratio in online learning: insights from MOOC". *Behaviour & Information Technology* 39.3, pp. 324–340 (cit. on p. 11).

Wang, Yurou, Haobo Zhang, Jue Wang, and Xiaofeng Ma (2023). "The Impact of Prompts and Feedback on the Performance during Multi-Session Self-Regulated Learning in the Hypermedia Environment". *Journal of Intelligence* 11.7, p. 131 (cit. on p. 126).

Wauters, Kelly, Piet Desmet, and Wim Van Den Noortgate (2011). "Monitoring Learners' Proficiency: Weight Adaptation in the Elo Rating System". *EDM*, pp. 247–252 (cit. on p. 66).

— (2012). “Item Difficulty Estimation: An Auspicious Collaboration Between Data and Judgment”. *Computers & Education* 58.4, pp. 1183–1193 (cit. on p. 38).

Wauters, Kelly, Piet Desmet, and Wim Van den Noortgate (2010). “Adaptive item-based learning environments based on the item response theory: Possibilities and challenges”. *Journal of Computer Assisted Learning* 26.6, pp. 549–562 (cit. on p. 65).

Westlin, Joseph, Eric Anthony Day, and Michael G Hughes (2019). “Learner-controlled practice difficulty and task exploration in an active-learning gaming environment”. *Simulation & Gaming* 50.6, pp. 812–831 (cit. on p. 58).

White, Kinnards (1968). “Delay of test information feedback and learning in a conventional classroom”. *Psychology in the Schools* 5.1, pp. 78–81 (cit. on p. 142).

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York (cit. on p. 68).

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller (2023). *dplyr: A Grammar of Data Manipulation* (cit. on p. 68).

Wilson, R. C., A. Shenhav, M. Straccia, and J. D. Cohen (2019). “The Eighty Five Percent Rule for Optimal Learning”. *Nature Communications* 10.1, p. 4646 (cit. on pp. 25, 59, 75, 173).

Winne, Philip H (2017). “Learning analytics for self-regulated learning”. *Handbook of learning analytics* 754, pp. 241–249 (cit. on p. 8).

Wintoro, Puput Budi and Mahendra Pratama (2022). “Course clustering in Moodle based learning management system using unsupervised learning”. *AIP Conference Proceedings*. Vol. 2563. 1. AIP Publishing (cit. on p. 8).

Wise, Alyssa Friend (2014). “Designing pedagogical interventions to support student use of learning analytics”. *Proceedings of the fourth international conference on learning analytics and knowledge*, pp. 203–211 (cit. on p. 175).

— (2019). “Learning analytics: Using data-informed decision-making to improve teaching and learning”. *Contemporary technologies in education: Maximizing student engagement, motivation, and learning*, pp. 119–143 (cit. on p. 8).

Wisniewski, Benedikt, Klaus Zierer, and John Hattie (2020). “The power of feedback revisited: A meta-analysis of educational feedback research”. *Frontiers in psychology* 10, p. 487662 (cit. on pp. 29, 30, 83, 112).

Wong, Sarah Shi Hui and Stephen Wee Hun Lim (2022). “Deliberate errors promote meaningful learning.” *Journal of Educational Psychology* 114.8, p. 1817 (cit. on p. 28).

World Bank (2024). *Impact Evaluations for Education Policy*. Tech. rep. World Bank (cit. on p. 2).

Xu, Mingfei and Simin Zeng (2023). “Optimal timing of treatment for errors in second language learning–A systematic review of corrective feedback timing”. *Frontiers in Psychology* 14, p. 1026174 (cit. on pp. 85, 86, 112, 116, 141, 171).

Xu, Zhihong, Kausalai Wijekumar, Gilbert Ramirez, Xueyan Hu, and Robin Irey (2019). “The effectiveness of intelligent tutoring systems on K-12 students’ reading comprehension: A meta-analysis”. *British Journal of Educational Technology* 50.6, pp. 3119–3137 (cit. on p. 7).

Yan, Veronica X, Yue Yu, Michael A Garcia, and Robert A Bjork (2014). “Why does guessing incorrectly enhance, rather than impair, retention?” *Memory & Cognition* 42, pp. 1373–1383 (cit. on p. 31).

Yaneva, Victoria, Peter Baldwin, Janet Mee, et al. (2019). “Predicting the difficulty of multiple choice questions in a high-stakes medical exam”. *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pp. 11–20 (cit. on p. 26).

Yaseen, Husam, Abdelaziz Saleh Mohammad, Najwa Ashal, Hesham Abusaimeh, Ahmad Ali, and Abdel-Aziz Ahmad Sharabati (2025). “The Impact of Adaptive Learning Technologies, Personalized Feedback, and Interactive AI Tools on Student Engagement: The Moderating Role of Digital Literacy”. *Sustainability* 17.3, p. 1133 (cit. on p. 6).

Yilmaz, Yucel and Ayşenur Sağdıç (2019). “The interaction between inhibitory control and corrective feedback timing”. *ITL-International Journal of Applied Linguistics* 170.2, pp. 204–227 (cit. on p. 126).

Yuhana, Umi Laili, Arif Djunaidy, Eric Pardede, Mauridhi Hery Purnomo, et al. (2024). “Clustering Approach for Modeling Course Difficulty Level in Adaptive Learning”. *2024 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. IEEE, pp. 1–6 (cit. on p. 25).

Zawadzka, Katarzyna, Oliwia Zaborowska, Ewa Butowska, Krzysztof Piątkowski, and Maciej Hanczakowski (2023). "Guessing can benefit memory for related word pairs even when feedback is delayed". *Memory & Cognition* 51.5, pp. 1235–1248 (cit. on p. 126).

Zhang, Ke and Ayse Begum Aslan (2021). "AI technologies for education: Recent research & future directions". *Computers and Education: Artificial Intelligence* 2, p. 100025 (cit. on p. 5).

Zhang, Qi and Zhonggen Yu (2022). "Meta-Analysis on Investigating and Comparing the Effects on Learning Achievement and Motivation for Gamification and Game-Based Learning". *Education Research International* 2022.1, p. 1519880 (cit. on p. 77).

Zhang, Qian and Logan Fiorella (2023). "An integrated model of learning from errors". *Educational Psychologist* 58.1, pp. 18–34 (cit. on pp. 28, 29).

Zhang, Yaqian and Wooi-Boon Goh (2021). "Personalized task difficulty adaptation based on reinforcement learning". *User Modeling and User-Adapted Interaction* 31.4, pp. 753–784 (cit. on pp. 27, 58).

Zheng, Lanqin, Jiayu Niu, and Lu Zhong (2022). "Effects of a learning analytics-based real-time feedback approach on knowledge elaboration, knowledge convergence, interactive relationships and group performance in CSCL". *British Journal of Educational Technology* 53.1, pp. 130–149 (cit. on pp. 6, 7).

Ziegler, Nicole (2016). "Synchronous computer-mediated communication and interaction: A meta-analysis". *Studies in Second Language Acquisition* 38.3, pp. 553–586 (cit. on p. 142).

Zou, Xiaotian, Wei Ma, Zhenjun Ma, and Ryan S Baker (2019). "Towards helping teachers select optimal content for students". *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20*. Springer, pp. 413–417 (cit. on pp. 57, 75).

RÉSUMÉ

L'apprentissage numérique offre une opportunité unique d'intégrer des stratégies fondées sur des preuves—telles que le feedback systématique, la pratique du rappel et l'ajustement adaptatif de la difficulté—tout en permettant un apprentissage personnalisé grâce au suivi précis des progrès individuels. Cette thèse examine l'optimisation de l'apprentissage numérique dans l'enseignement médical en adoptant une perspective en sciences cognitives. Plus précisément, elle aborde deux questions fondamentales : l'optimisation de la difficulté des entraînements et l'utilisation efficace du feedback pour améliorer les performances d'apprentissage. Pour y répondre, une approche multi-méthodes a été adoptée, combinant une méta-analyse systématique, l'analyse de l'apprentissage (learning analytics) et une expérimentation à grande échelle sur UNESS-BNE, un système numérique largement utilisé par les étudiants en médecine en France.

Le chapitre 1 introduit le contexte général et la motivation de cette thèse. Le chapitre 2 porte sur l'adaptation du système de notation Elo, couramment utilisé, pour estimer à la fois les compétences des étudiants et la difficulté des questions sur la plateforme UNESS-BNE. Les résultats montrent que le système Elo atteint une performance prédictive comparable à celle des modèles logistiques bien calibrés pour prédire les résultats finaux aux examens, confirmant ainsi sa pertinence pour ce jeu de données. Sur cette base, le chapitre 3 exploite le modèle adapté au sein d'UNESS-BNE afin d'examiner le niveau optimal de difficulté d'entraînement dans les questions à choix multiples en formation médicale. Les résultats soutiennent l'hypothèse en U inversé, indiquant un niveau optimal de difficulté dans ce contexte d'apprentissage.

Le chapitre 4 présente une méta-analyse des effets du moment du feedback dans les environnements numériques. Les résultats indiquent que, globalement, le timing du feedback n'influence pas significativement les performances d'apprentissage. Cependant, les analyses de modération révèlent l'impact de facteurs tels que le niveau éducatif, le domaine d'apprentissage, le type de tâche lors du post-test et les contraintes temporelles de réponse, offrant une explication partielle aux incohérences observées dans les études antérieures. Enfin, le chapitre 5 présente une étude expérimentale portant sur les effets individuels et interactifs du moment du feedback (immédiat vs différé) et du rappel initial de la réponse sur les performances d'apprentissage en formation médicale. Bien qu'aucun effet significatif n'ait été mis en évidence, vraisemblablement en raison d'un engagement limité des participants et d'une exposition insuffisante à la manipulation expérimentale, la collecte de données se poursuit et pourrait permettre d'obtenir des résultats plus concluants à mesure qu'elle progresse.

Ensemble, ces études offrent des perspectives précieuses sur l'optimisation de l'apprentissage numérique dans l'enseignement médical. Les résultats apportent des implications pratiques pour les praticiens et les acteurs de l'éducation, tout en contribuant à une meilleure compréhension des processus d'apprentissage, de la mémoire et de la recherche éducative dans les environnements numériques.

MOTS CLÉS

Apprentissage numérique, Éducation médicale, Analyse de l'apprentissage

ABSTRACT

Digital learning presents a unique opportunity to incorporate evidence-based strategies—such as systematic feedback, retrieval practice, and adaptive difficulty adjustment—while enabling personalized learning through precise monitoring of individual progress. This dissertation investigates the optimization of digital learning within medical education, adopting a cognitive science perspective. Specifically, this research addresses two fundamental questions: the optimization of training difficulty and the effective use of feedback to enhance learning outcomes. To answer these questions, a multi-method approach was employed, combining systematic meta-analysis, learning analytics, and large-scale experimentation on UNESS-BNE, a widely used digital learning system for French medical students.

Chapter 1 introduces the general context and motivation of the dissertation. Chapter 2 focuses on adapting the widely used Elo rating system to estimate both student ability and question difficulty in the UNESS-BNE platform. The results demonstrate that the Elo rating system achieves predictive performance comparable to well-calibrated logistic models in predicting students' final exam outcomes, confirming its suitability for this dataset. Building on this, Chapter 3 utilizes the adapted model within the UNESS-BNE system to examine the optimal level of training difficulty in the context of multiple-choice questions in medical education. The findings support the inverted U-shaped hypothesis, indicating the presence of an optimal difficulty level in this specific learning setting.

Chapter 4 presents a meta-analysis of the effects of feedback timing in digital learning environments. The results indicate that, overall, feedback timing does not significantly influence learning outcomes. However, moderator analyses highlight the impact of factors such as educational level, learning domain, post-test task type, and response time constraints, providing a partial explanation for inconsistencies observed across previous studies. Finally, Chapter 5 details an experimental study investigating the individual and interactive effects of feedback timing (immediate vs. delayed) and initial answer recall on learning outcomes in medical training. Although no significant effects were found, likely due to limited engagement and exposure to the manipulation, data collection is ongoing and may support more conclusive results as it progresses.

Taken together, these studies provide valuable insight into optimizing digital learning in medical education. The findings offer practical implications for practitioners and stakeholders in the broader education sphere, contributing to our understanding of learning, memory, and educational research in digital environments.

KEYWORDS

Digital Learning, Medical Education, Learning Analytics