UNIVERSITY OF CALIFORNIA,
IRVINE

Individual Differences in Working Memory and Mathematical Skill Learning

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Yi Feng

Dissertation Committee:
Professor Susanne M. Jaeggi, Chair
Professor Mark Steyvers
Assistant Professor Shayan Doroudi

2024

© 2024 Yi Feng

DEDICATION

This dissertation is dedicated to my beloved parents, who have always been my strongest supporters. As their only child, they have allowed me to pursue my dreams and ambitions, even when it meant I couldn't be by their side. Their unwavering support, encouragement, and constant wish for my health and happiness have been my foundation. Without them, I could not have completed this journey.

I also want to express my deepest gratitude to my dear friends who have accompanied me along this path. To Sirui, who served as the bridge connecting me to UCI and my advisor, Susanne M. Jaeggi. Your friendship, support, and the sense of belonging and security you've provided have been invaluable. To Xuehan Zhou and Yijun Xiang, your friendship brought me joy and made this journey a wonderful experience. To Yannan Zhou and Jiali Wang, your support, especially during my most emotional and lonely times, gave me the strength to keep going.

A special thanks to Xunfei Li, whose kindness, strength, and empathy have been a guiding light. You have taught me to be brave and to embrace everything life has to offer. Without the support of all my friends, I wouldn't have come this far.

I also want to extend my thanks to Nikki, Cooper, and Aja, who filled my days with happiness and were the best exercise companions. I cherish the bonds we've formed and will keep in touch with all of you.

Finally, I want to thank my boyfriend, who helped me navigate this journey with his coding skills and unwavering support through many life events. Thank you for being there every step of the way.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	v
ACKNOWLEDGEMENTS	vi
VITA	vii
ABSTRACT OF THE DISSERTATION	viii
INTRODUCTION	1
CHAPTER 1: Introduction	10
Methods	16
Results	33
Discussion	40
Supplementary Materials	48
CHAPTER 2: Introduction	60
Methods	64
Results	70
Discussion	82
Supplementary Materials	87
CHAPTER 3: Introduction	89
Methods	93
Results	101
Discussion	110
DISCUSSION	117
REFERENCES	124

LIST OF FIGURES

		Page
Figure 1.1	General Flowchart of Analytic Plan	32
Figure 1.2	Learning Patterns Identified by Clustering	34
Figure 1.3	Predictor Importance by Calculating the SHAP Value	38
Figure 1.4	Scatter Plots of Predictor Values (x axis) and SHAP Values (y axis)	40
Figure 2.1	Experimental design across the three studies	67
Figure 2.2	Net effect sizes of training groups in each study	75
Figure 2.3	WM and math improvements as a function of baseline math skills and training types	82
Figure 3.1	Three Executive Function + Math Games used in Training Program	96
Figure 3.2	Aggregated Learning Trajectory	96
Figure 3.3	Training Procedure	98
Figure 3.4	Math Measures	100
Figure 3.5	Working Memory Measure	101
Figure 3.6	Training Effects on Working Memory and Math Performance	102
Figure 3.7	Models' Performance in Predicting Training Improvement	106
Figure 3.8	Clustering Goodness Evaluation	107
Figure 3.9	Clustering Results Based on First Three Sessions Performance	108
Figure 3.10	Comparisons Between Aloe and Bamboo Groups	109
Figure 3.11	Subgroup Comparisons Within the Aloe Group	110
Figure 3.12	Prediction Error Distribution	114

LIST OF TABLES

		Page
Table 1.1	Confusion Matrix of Binary Tree Model Performance on Cross Validation Set	36
Table 1.2	Confusion Matrix of Binary Tree Model Performance on Test Set	37
Table 2.1	Descriptive statistics of demographic information across the three studies	65
Table 2.2	Regression results for baseline performance	72
Table 2.3	Regression results for mathematics outcomes	78
Table 2.4	Regression results for WM outcomes	80

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my committee chair, Professor Susanne M. Jaeggi, who has been the most exceptional advisor I could have hoped for. Her kindness, empathy, and understanding have been like a guiding light, illuminating the path for me, especially as a non-native English speaker and international student. She has not only helped me navigate the complexities of research but has also taught me how to communicate effectively and organize my studies. Professor Jaeggi has always gone above and beyond, connecting me with other experts and supporting my growth in every way possible. I feel incredibly fortunate to have her as my advisor.

I also want to extend my gratitude to her husband, Martin Buschkuehl, who has been both a great collaborator and a nice coach during my research experience. Together, they have provided me with a positive research environment and a strong sense of belonging. Without their guidance and support, this dissertation would not have been possible. I am profoundly grateful to both of them.

I would also like to thank my committee members, Professor Mark Steyvers and Professor Shayan Doroudi. Their expertise and insightful suggestions have greatly enriched my dissertation. I appreciate their willingness to serve on my committee and their valuable contributions during both my proposal and dissertation defenses.

In addition, I am deeply thankful to my collaborators, Professor Geetha B. Ramani, Professor Aaron R. Seitz, and Professor Dennis Barbour. Their guidance and instruction have helped me grow immensely, and their expertise has allowed me to participate in some truly wonderful projects. I have learned so much from each of them, and I am incredibly grateful for their support.

A special thanks to Anja, Dom, Masha, Ally, Mary, and all the other lab members. Their support, kindness, and willingness to share their expertise and ideas have been invaluable. They have helped me in several projects, and we have had some truly great times together. I genuinely appreciate all of them.

VITA

Yi Feng

2015 B.S. in Psychology, South China Normal University

2018 M.S. in Cognitive Science, Beijing Normal University

2018-2019 Research Assistant, Beijing Normal University

2024 Ph.D. in Education, University of California, Irvine

FIELD OF STUDY

Human Learning, Cognitive Training, Individual Differences in Learning

PUBLICATIONS

DePascale, M., **Feng, Y.,** Lin G. C., Barkin R., Akhavein K., Tavassolie N., Ghil E., Gaye F., Buschkuehl M., Ramani1, G. B., & Jaeggi, S. M. (2023). Uncovering the reciprocal relationship between domain-specific and domain-general skills: Combined numerical and working memory training improves children's mathematical knowledge. Contemporary Educational Psychology, 102252. (IF: 10.3)

Feng, Y., Pahor, A., Seitz, A. R., Barbour, D. L., & Jaeggi, S. M. (2023). Unicorn, Hare, or Tortoise? Using Machine Learning to Predict Working Memory Training Performance. Journal of Cognition, 6(1).

Tullo, D., **Feng, Y.,** Pahor, A., Cote, J. M., Seitz, A. R., & Jaeggi, S. M. (2023). Investigating the Role of Individual Differences in Adherence to Cognitive Training. Journal of Cognition, 6(1).

Zhao, H., Li, X., Karolis, V., **Feng, Y.,** Niu, H., Butterworth, B. (2019).Arithmetic learning modifies the functional connectivity of the fronto-parietal network. Cortex, 51-62(111).

Lu, A., Tian, H., Yu, Y., **Feng, Y.,** Hong, X., & Yu, Z. (2015). Peer attachment and social anxiety: Gender as a moderator across deaf and hearing adolescents. Social Behavior and Personality: an international journal, 43(2), 231-239.

Lu, A., Yu, Y., Hong, X., **Feng, Y.,** Tian, H., & Liao, J. (2014). Peer attachment and loneliness among adolescents who are deaf: The moderating effect of personality. Social Behavior and Personality: an international journal, 42(4), 551-560.

Lu, A., **Feng, Y.,** Yu, Z., Tian, H., Hong, X., & Zheng, D. (2015). Anxiety and mind wandering as independent consequences of stereotype threat. Social Behavior & Personality An International Journal, 43(4).

ABSTRACT OF THE DISSERTATION

Individual Differences in Working Memory and Mathematical Skill Learning

by

Yi Feng

Doctor of Philosophy in Education

University of California, Irvine, 2024

Professor Susanne M. Jaeggi, Chair

Individual differences significantly impact learning outcomes, yet research has been limited in predicting training outcomes based on these differences and in examining how they interact with intervention types. Understanding these dynamics is crucial for developing effective training programs. This dissertation addressed these gaps through three studies.

Study 1 examined how individual differences predict variations in learning trajectories during working memory training in undergraduates. Participants were clustered into three training patterns, and machine learning models predicted these patterns based on baseline cognitive abilities, personal characteristics, environmental and experience-related factors. Openness and pre-existing working memory capacity were key predictors of high performance, while openness and video game background predicted learning persistence among lower performers. This study demonstrates the potential to predict individual training performance using participant characteristics before training.

Study 2 explored how baseline abilities interacted with three interventions: working memory-only, numerical knowledge-only, and combined working memory + numerical

knowledge training in kindergarteners. Overall, the study found that combined training led to the most pronounced long-term gains in math skills. Moreover, participants with high initial math skills showed the greatest improvements in both working memory and math with combined training, whereas numerical knowledge-only training was more effective for children with lower initial math skills. This suggests that combined training is generally efficient, but baseline abilities also matter.

Study 3 built on Studies 1 and 2 by integrating executive function training within a math invention for children. Using early learning data to predict training improvement, the study found that performance in the first three sessions could predict overall training improvement and identified two learning patterns: a group with higher initial scores and greater improvement, and a group with lower initial scores and slower improvement. Within the lower improvement group, some participants still showed significant progress, catching up to the high improvement group, likely due to completing more training sessions.

Overall, these studies highlight the importance of considering individual differences and the type of cognitive intervention in designing optimally tailored programs. They also illustrate the potential of using pre-training or early individual differences in learning trajectories to predict training outcomes, which has implications for personalized educational strategies.

INTRODUCTION

Background

Can all students acquire all skills equally well? Why are some students better able to acquire certain skills over others? Different individuals learn differently, which has long been a concern for classroom teachers. Teachers do not have much choice in this regard: students have wide variations in learning rates and styles, while teachers have limited tools and time to meet these challenges. Too often, it is a lack of information and resources rather than neglect that leads to students' needs not being met. The structure of group instruction in schooling already makes it challenging for teachers to address every student's individual needs. On top of that, teachers often don't have the information or knowledge regarding how and why individuals differ in their learning abilities. Meeting individual needs in a large group setting is a pervasive challenge that needs to be addressed.

Researchers play a crucial role in helping solve this problem. To provide more information and resources, researchers need to determine the extent of individual differences in skill learning, identify the factors causing these differences, and offer evidence-based and targeted interventions that can be implemented beyond typical classroom settings. Traditionally, individual differences in learning were often overlooked and treated as error terms in psychological and educational experiments (e.g., Fleishman, 1967; Jensen, 1967). This oversight led to misleading theories of learning and unexplained discrepancies among studies (Farr, 1971). However, accumulating research has started to shed light on individual differences in skill learning. Foundational work by Cronbach &

Snow (1969) and Ackerman (1987, 1988) underscores the importance of recognizing and explaining individual differences in learning. Their research quantified how these individual differences evolve over time and observed that different teaching methods are effective for different types of learners. Numerous subsequent studies have explored these topics, identifying various reasons for differing individual learning styles. Factors such as aptitudes for learning, willingness to learn, and learning preferences impact the learning process for each student (Sauce & Matzel, 2013). These learner traits have shown to determine at least to some degree how well any individual can learn (Ackerman et al., 2001; Katz et al., 2021; Smid et al., 2020).

Given the challenge of providing personalized instruction in schools, targeted cognitive training offers a solution to supplement classroom learning. Early intervention studies aimed to improve specific skills by repeatedly practicing domain-specific tasks, such as motor training and mathematical training (Bekoff, 1988; Starkey et al., 2004; Whitehurst et al., 1994). Later, researchers hypothesized that interventions targeting domain-general skills, such as executive functions, might benefit a wide range of specific skills, including reading, mathematics, and problem-solving skills (Hambrick & Engle, 2003; McVay & Kane, 2012; Raghubar et al., 2010; Swanson & Beebe-Frankenberger, 2004). Executive function (EF) is a set of cognitive processes essential for learning behavior, including working memory, inhibitory control, and cognitive flexibility. Working memory (WM) refers to the ability to temporarily hold and manipulate information for a given task (Cowan, 2008; Miyake & Shah, 1999); inhibitory control refers to the capacity to inhibit unrelated impulses and focus attention on current tasks (Williams et al., 1999); cognitive flexibility refers to the ability to shift cognitive status to adapt to different tasks (Scott,

1962; Boger-Mehall, 1996). To illustrate the importance of EFs, consider a student's daily activities, such as focusing during lessons, shifting between different subjects, and organizing assignments. EFs help students maintain concentration, adapt to changing tasks, and manage their workload effectively. Therefore, many researchers have been interested in improving EF skills, demonstrating that these skills can be enhanced through targeted training (Au et al., 2016; Jaeggi et al., 2011, 2008; Pahor et al., 2022; Von Bastian et al., 2013). However, some studies do not find EF training to be effective (Melby-Lervåg & Hulme, 2016; Thompson et al., 2013). Some researchers argue that individual differences might explain these mixed results (Katz et al., 2021; Meiran et al., 2019; Redick, 2019; Wiemers et al., 2019). This suggests that some individuals may not benefit from training, and training may even increase differences between them. For example, Ackerman (2007) demonstrates that practicing a complex task accentuates existing differences between individuals, suggesting that the training process contributes to these differences. The cognitive training field faces challenges due to uncertainty regarding training effects. If we do not understand the mechanisms underlying individual differences and find appropriate training methods for different types of students, cognitive training may not be a reliable supplement to education in its current form.

In developing optimal intervention programs, it is also important to consider the interaction between individual attributes and the intervention. Several theories suggest that individuals benefit when learning materials or training programs are tailored to their specific attributes. For example, Vygotsky (1978) developed the concept of the "Zone of Proximal Development," proposing that students improve more with appropriate guidance than without it, with the range of growth depending on their basic knowledge. Each

student's zone of proximal development is different, suggesting that teachers need to plan targeted instruction to match personal needs. Similarly, the Aptitude by Treatment Interaction (ATI) theory (Snow, 1989; Cronbach & Snow, 1977) argues that optimal training gains result from precisely matched interventions to individual aptitudes. It highlights aptitudes' moderation effect on training outcomes, potentially explaining variability in interventions. As a result, training outcomes might be predicted based on the combination of personal aptitudes and certain training programs.

In relation to ATI, there are two main goals to realize personalized learning. First, personality dimensions and aptitudes can serve as criteria for program placement. However, personal profiles are often complex, making it challenging to match interventions based on individual attributes because it is difficult to determine which factors should be emphasized. Second, different interventions should be employed for different individuals to achieve the same learning goals, ensuring that each student's unique needs and strengths are addressed. To achieve these goals, we need to investigate the following: i) What individual attributes determine learning effects, and to what extent? ii) Is there an optimal intervention among different interventions, and how do individual attributes interact with different interventions? iii) How can individual differences predict training outcomes?

Research on these aspects arrives at a perfect time due to advances in computer technology. Cognitive training is now accessible to a wider audience due to the ubiquity of mobile devices. We can use computerized tasks or games for both training and assessments, enabling us to provide personalized training by i) conveniently collecting individual attributes before training, ii) recognizing that each individual can complete

different tasks via personal devices, and iii) tracking individual progress during training. This allows us to gather students' learning data in real-time, enabling us to go beyond predicting individual differences after training. We can explore predicting training outcomes during intervention, allowing us to adjust interventions as needed to maximize individual progress.

With facing both challenges and opportunities in the area of cognitive training, there are several research gaps. Regarding individual differences in learning, numerous factors influence how much an individual learns from training, including baseline abilities, age, motivation, and personality (see review by Katz et al., 2021). The mechanism by which individual differences influence training outcomes is complex. We still know little about the determinant factors and the extent to which they predict training outcomes. Regarding the interaction between individual attributes and interventions, specifically in math learning, previous research has tested the benefits of different types of intervention (Bergman-Nutley & Klingberg, 2014; DeWind & Brannon, 2012; Park et al., 2016; Ribner, 2020; Sánchez-Pérez et al., 2018; Söderqvist & Bergman, 2015), but it is unclear how individual attributes interact with domain-specific and domain-general interventions. Lastly, previous research has rarely attempted to predict training outcomes (Bürki et al., 2014; Guye et al., 2017; Karbach et al., 2017). Focusing on the predictive capabilities of individual difference factors could help uncover the intricate relationships between multiple factors and learning outcomes, and forecast future behavior (Luan & Tsai, 2021; Orrù et al., 2020; Yarkoni & Westfall, 2017). Additionally, early prediction of individual training outcomes can help adjust interventions at the right time, but research in this area has been scant.

To fill these research gaps, this dissertation focused on uncovering the role of individual differences in cognitive training through three studies. The aim of Study 1 was to determine which individual difference factors predict variations in cognitive training and how. Using machine learning approaches, I explored young adults' variability in working memory skill learning and examined how baseline cognitive abilities, personal characteristics, environmental, and experience-related factors predict this variability. Studies 2 and 3 concentrated on mathematics learning, a challenging area for many children. Study 2 examined the effects of three different training methods (domain-general skills training only, domain-specific skills training only, and both) and the interaction between individual attributes and different treatments on training outcomes. Study 3 tested the benefits of newly developed cognitive training games that combined domain-general and domain-specific skills to improve mathematical skills in a classroom setting. Building on Study 2, it investigated the efficacy of integrating domain-specific and domain-general skills training and further tested early prediction of training outcomes.

Significance

This dissertation bolsters scholarly and public understanding of personalized education. Considering that individuals have differing abilities, it starts from the premise that the best way to accommodate these differences is by identifying and grouping individuals according to their strengths and needs. By doing so, students can be taught target skills that will benefit them most, thereby maximizing their learning outcomes.

Specifically, Study 1 examined the working memory skill learning trajectories of undergraduates, grouping participants based on these trajectories, identifying different learning patterns, and using several individual variables in machine learning models to

predict these patterns. The study provided evidence that learning trajectories can indeed be predicted and identified key factors such as openness and pre-existing working memory capacity as predictors of high performance, while openness and video game background predicted learning persistence among lower performers. These factors can give practitioners insights into developing different training programs to meet the needs of diverse populations. Additionally, the study explored the possibility of using a prediction model to forecast learning patterns before training.

Study 2 shifted the focus by examining the interaction between several treatments and children's initial skills. This study provided insight into the most effective treatment for improving the mathematical skills of children. Mathematical skills are critical for academic achievement and future STEM career development (Blotnicky et al., 2018). Several approaches have been implemented to promote individuals' math performance. One approach emphasized designing domain-specific interventions to directly target math skills (e.g., Clements & Sarama, 2008; Jordan et al., 2012; Siegler & Ramani, 2008, 2009), while another focused on training individuals' domain-general skills, such as executive functions (EFs)(Kroesbergen et al., 2012, 2014; Titz & Karbach, 2014; Wang et al., 2019), or combined these two approaches with the idea that this could maximize the training benefits (DePascale et al., 2024; Fuchs et al., 2020; Lee, 2023; Nemmi et al., 2016; Prager et al., 2023; Ribner and Libertus, 2021). This dissertation expanded on these approaches by considering how learner attributes impact the outcomes of learners in each possible treatment, and which treatments might be most appropriate for particular learners. In Study 2, several treatments for improving mathematical skills were investigated, and individuals' initial skills were used as individual difference variables to see how they

interacted with treatments. We found that the intervention incorporating both domain-general and domain-specific training generally yielded the most substantial improvements in math learning. It is also noteworthy that for both math and working memory outcomes, individuals with lower initial math skills tended to benefit more from single-type (domain-specific) training, whereas those with higher initial skills benefited from combined training. The study found that the intervention incorporating both domain-general and domain-specific training generally yielded the most substantial improvements in math learning. Notably, individuals with lower initial math skills tended to benefit more from single-type (domain-specific) training, whereas those with higher initial skills benefited more from combined training. This finding underscores the importance of tailoring interventions to individual learner profiles to maximize educational outcomes.

Study 3 built on lessons learned from Studies 1 and 2, developing training interventions for mathematical learning that are broadly effective for individuals with a variety of baseline abilities. Besides examining the efficacy of the training program, the primary goal was to take a further step by monitoring students' training gains through predictive models. Specifically, a prediction model aimed to predict individuals' training gains as early as possible. Early prediction can facilitate timely interventions, allowing for personalized adjustments that optimize the learning experience for each participant. This approach is crucial for identifying individuals who may struggle and need additional support to fully benefit from the training program, allowing for timely adjustments and additional assistance early in the intervention process. Study 3 found that performance data from the first three sessions could be used to forecast training improvement across the entire period using a linear regression algorithm. With just the first three sessions, we

were able to group participants into two distinct groups. This identification point offers a critical timing for adjustments, such as gathering feedback or implementing motivational techniques, to optimize training outcomes.

CHAPTER 1

Unicorn, Hare, or Tortoise? Using Machine Learning To Predict Working Memory Training Performance

Published in: Feng, Y., Pahor, A., Seitz, A. R., Barbour, D. L., & Jaeggi, S. M. (2023). Unicorn, Hare, or Tortoise? Using Machine Learning to Predict Working Memory Training Performance. *Journal of Cognition*, 6(1).

1.1 Introduction

If you happen to have ever been on an athletic team, you probably noticed that even when team members are given the same training program, their progress varies greatly. Some athletes respond well and show huge improvements at the beginning of the training program. Others improve very slowly, even if they practice the same amount of time as those who show rapid improvements. It is undeniable that there is considerable variation between people during training and the extent to which they learn and improve their performance. The question remains, however, as to what factors are responsible for these inter-individual variations in learning adaptation. In this study, we explore the variability in learning to answer this question. We use working memory (WM) training as a proxy for learning since WM - the cognitive process that facilitates temporarily holding and manipulating information for a short period of time - is a good indicator of general skill learning ability (Cowan, 2008).

Heterogeneity of Learning Trajectories

There is an ongoing debate around the universal effectiveness of WM training (Au et al., 2016; Jaeggi et al., 2008, 2011; Klingberg, 2010; Melby-Lervåg & Hulme, 2016; Pahor,

Seitz, & Jaeggi, 2022; Thompson et al., 2013; Von Bastian et al., 2013), and it has been argued that a major source of variability can be attributed to individual differences (Borella et al., 2017; Jaeggi et al., 2014; Traut, Guild, & Munakata, 2021). This heterogeneity in WM training is evident in both training outcomes (Katz et al., 2021; Meiran et al., 2019; Redick, 2019; Wiemers et al., 2019), as well as in learning trajectories that emerge during the intervention process (Flak et al., 2019; Holmes et al., 2009; Karbach et al., 2015). Additionally, Ackerman (2007) demonstrated that practicing a complex task accentuates existing individual differences, implying that the training process itself plays a role in shaping these variations. Notably, learning trajectories vary significantly between individuals, even when performing the same task (Bürki et al., 2014; Guye et al., 2017; Wiemers et al., 2019), and have shown to predict learning outcomes (Karbach, Strobach, & Schubert, 2015; Jaeggi et al., 2011), however, individual differences in learning trajectories within the cognitive training literature remain relatively sparse (e.g., Bürki et al., 2014; Guye et al, 2017; Karbach et al., 2017; Matysiak, Kroemeke, & Brzezicka, 2019, Ørskov et al., 2021).

Könen and Karbach (2015) emphasized the need for research that includes comprehensive training trajectory data to better understand the factors contributing to individual differences in skill learning. They suggested that analyzing time-series training data can provide valuable insights, such as identifying sessions where participants improved, their peak performance, the stability or variation between sessions, and the turning points when stability changed. Accordingly, these findings will provide insights on how to design customized training settings for individuals based on their learning curves. However, only a few studies have examined individual differences using dynamic training

data (Bürki et al., 2014; Guye et al., 2017; Karbach et al., 2017). These studies have typically employed latent growth models to analyze training trajectories, assuming linear or quadratic changes in performance. However, these models have limitations, as they assume predefined trajectories applicable to all participants, and stable performance changes throughout the intervention period. Thus, methods to capture individual learning trajectories beyond these indicators are needed, for example by including when and how stability has changed, which is the goal of the present paper.

Factors Affecting Skill Learning

Several factors have been identified as determinants of skill learning. These factors encompass both pre-existing cognitive ability and non-cognitive traits such as interests, personality, motivation, and self-efficacy, all of which have shown to influence skill acquisition (Ackerman, 2007; Bastian & Oberauer, 2014; Katz et al., 2021). However, no specific theory has unified all these findings so far. Bandura's social-cognitive framework, which considers interactions between behavioral, person-centered, and environmental factors, offers a comprehensive perspective on individual learning and performance (Bandura, 1986). In this study, we rely on Bandura's framework and focus on prominent factors that have shown to impact learning outcomes in previous studies within the WM training literature, categorizing them into three groups: baseline cognitive ability, person-related characteristics, as well as environmental and experience-related factors.

Baseline cognitive ability, which refers to pre-existing ability before any intervention, significantly influences skill learning (Bürki et al., 2014; Foster et al., 2017; Guye et al., 2017; Karbach et al., 2017; Zinke et al., 2012, 2014; Ørskov et al., 2021). Various measures of baseline cognitive ability have been found to impact skill learning outcomes,

which include general cognitive skills (such as fluid reasoning), as well as specific tasks that share processes with trained tasks. However, there are still uncertainties regarding the relationship between baseline cognitive ability and learning performance. Some studies suggest that individuals with the highest initial performance on trained tasks show the most improvement during WM training (magnification account or Matthew's effect; "rich-get-richer"; e.g., Bürki et al., 2014; Foster et al., 2017; Guye et al., 2017; Ørskov et al., 2021). Conversely, other research indicates that individuals with the lowest initial training performance experience the largest gains in WM training (compensation account; "catch up"; e.g., Jaeggi et al., 2011; Karbach et al., 2017; Zinke et al., 2012, 2014). Additionally, learning outcomes might be differentially impacted by general cognitive ability or specific skills (Karbach et al., 2017; Wiemers et al., 2019). To shed more light on this issue, our study aims to examine the impact of baseline performance by using general cognitive ability (measured by fluid reasoning), composite WM skills, as well as specific tasks sharing processes with trained tasks as predictors, in order to determine their impact on training results.

Person-related characteristics, including motivational factors, such as growth mindset or self-perceived cognitive failures, as well as certain personality traits have also shown to influence skill learning (Guye et al., 2017; Katz et al., 2021; Ørskov et al., 2021). For example, individuals with growth mindsets view training as a means to enhance their abilities, leading to increased engagement and improved performance (Dweck, 2000). Moreover, those reporting more cognitive failures have shown to be more likely to participate in a WM intervention and persist with them (Jaeggi et al., 2014). Personality traits like conscientiousness and neuroticism have also been linked to cognitive and

emotional regulation, aiding individuals in maintaining diligence and managing negative emotions during training (see review Katz et al., 2021). Other personality traits may also play a role in skill learning. For example, individuals who show high grit and ambition tend to persevere despite setbacks and hold high expectations for their performance, leading to increased engagement (Datu, 2021; Duckworth et al., 2007). However, the relationship between these person-related characteristics and learning outcomes has been inconsistent (Guye et al., 2017; Nemmi et al., 2016; Thompson et al., 2013). The scarcity of studies incorporating these person-related factors and the small sample sizes used in most studies present further challenges, which we aim to address in our study.

Environmental and experience-related factors also contribute to skill learning (Sigmundsson et al., 2017). Maintaining good physical health is crucial for optimal brain function and learning across the lifespan (Khan & Hillman, 2014; Macpherson et al., 2017). Although the direct impact of health on skill learning is not extensively studied, several studies have shown that physical exercise benefits spatial memory, working memory, and executive attention (see review Cassilhas et al., 2016). Furthermore, prior experience and exposure to cognitive training regimens or computerized tasks can promote skill learning (Berard et al., 2015; Steyvers & Schafer, 2020). Socioeconomic status (SES) has shown to be an indicator of individuals' exposure to experiences relevant to cognitive training including WM training, such as cognitive tutoring, but also playing computer games or educational puzzles (Duncan & Magnuson, 2012). Moreover, experience with computers and playing video games has been found to predict performance in cognitive training games, suggesting that gaming experience can enhance perceptual processing and learning more generally (Steyvers & Schafer, 2020; Zhang et al., 2021). This growing literature highlights the

potential importance of SES and video game experience in learning outcomes, but it warrants further exploration.

Predicting Trajectories

Previous research has examined individual difference variables to predict learning trajectories separately. However, skill learning is influenced by a combination of factors, and these factors are often interrelated, and the relationship might not be linear (e.g. Guye et al., 2017). Moreover, earlier studies have used limited approaches to examine links between individual difference factors and training performance. Even when using multiple regression or latent growth models, the assumed relationship between factors and training performance remains linear (Bürki et al., 2014; Guye et al, 2017; Karbach et al., 2017), which inherently restricts the interpretation of these findings. Even structural equation modeling, which uncovers complex relationships among variables, primarily relies on linear relationships and assumes independence among moderators. Research has revealed that when analyzing the impact of personality on cognitive ability through regression analysis, quadratic associations account for a significant amount of variance beyond the effects of linear factors (Major, Johnson & Deary, 2014). Thus, to comprehensively understand the relationship between individual difference factors and learning performance, it is necessary to employ an approach that can combine several factors and enable correlations between factors without assuming a linear relationship with training performance. Furthermore, in order to better understand underlying mechanisms of learning trajectories, focusing on the predictive capabilities of individual difference factors could help uncover the intricate and complex relationships between multiple factors and

learning outcomes and forecast future behavior (Luan & Tsai, 2021; Orrù et al., 2020; Yarkoni & Westfall, 2017).

Machine learning algorithms serve as the most advanced tools for prediction, as they can handle numerous predictors and their intricate relationships (Yarkoni & Westfall, 2017). Unlike traditional statistical explanatory methods that rely on assumptions about variable relationships, many machine learning algorithms employ nonlinear algorithms and do not make such assumptions. Specifically, the fundamental principle behind machine learning models is that they learn the relationships between variables in the provided data, enabling the prediction of future behaviors of new samples (Mohri et al., 2018).

In this study, we utilize machine learning models to predict learning trajectories in a WM intervention conducted by more than 500 participants (a large sample size in the cognitive training field). To do so, we rely on a set of predictors that have been identified as relevant in previous literature including baseline cognitive abilities, person-related characteristics, and environmental factors. We hypothesize that individuals' learning trajectories can be categorized into different patterns, and that a combination of individual difference variables can successfully predict these patterns using machine learning models. Given that many studies emphasize the role of baseline cognitive abilities in influencing training gains, we hypothesize that this factor will be the most powerful predictor.

1.2 Methods

Participants

Participants were recruited from the Universities of California, Irvine and Riverside. In total, 568 undergraduates from diverse socioeconomic backgrounds completed the WM intervention as well as several surveys and assessments. The data are combined from three

studies. Study 1 was conducted from 2017 to 2018 (266 participants); Study 2 from May to August 2020 (138 participants); Study 3 from May 2020 to December 2021 (164 participants). Study 1 was a pre-covid in-lab study, while Studies 2 and 3 were administered online due to the pandemic. Participants' age ranged between 18-57 (Mage = 20.37, SD=3.63), and 64% of them were women. They were diverse in terms of ethnicity and race, with 33% identifying as Hispanic or Latino, 3% identifying as African American, 46% Asian, 14% Caucasian, 0.7% American Indian, 8% Biracial/Mixed race, 0.9% Native Hawaiian or Other Pacific Islander, and 24% reporting others. The study procedures were approved by Institutional Review Boards at both sites, and each participant received $80 or $120 as financial compensation for the online or in-lab studies, respectively.

Training program

An app-based N-back training program was developed in-house by University of California Riverside Brain Game Center ("Recollect the Study"; available on Google Play and Apple App Store; cf. also https://apps.apple.com/us/app/recollect-the-study/id1217528682; a video can be found here: https://www.youtube.com/watch?v=zhgL8Oe42Yk). In the game, participants were presented with a series of stimuli consisting of different shapes and colors (see Supplementary Materials Figure S1). They were asked to respond to stimuli that matched the stimulus presented N items back. For instance, in a 1-back task, participants respond if the current stimulus is the same as the stimulus presented 1 item before. The task difficulty was adaptively adjusted based on participants' performance, with higher levels of N increasing the task difficulty. Participants need to keep track of more items with higher N levels. With the same task requirement, we employed different training environments,

including gamified and non-gamified variants, along with 10 adaptive algorithms to adjust the N-level progression (cf. Sandeep et al., 2020). For example, one algorithm required 3 hits to level up and 2 errors to level down within a block, while other algorithms utilized different conditions, such as 3 hits to level up and 3 errors to level down. For the purpose of the present work, our objective is to identify the influential factors that contribute to learning trajectories, irrespective of the specific training conditions. Therefore, we did not include the training condition as a predictor variable. Still, to explore whether training conditions would contribute variance to training trajectory, we added these variables to our model as described in the supplementary materials. Notably, the model performance did not change substantially (Table S7).

Each participant completed one 40-minute training session per day, with a break after 20 minutes, for a total of 10 sessions. Each session consisted of multiple blocks, and each block lasted 2-3 minutes that included 20 plus N trials. The dependent variable was the weighted average N level achieved during a session. Specifically, each N-level was multiplied by the number of trials associated with that block, then divided by the total number of trials. It represents the average level that participants achieved in a particular session (cf. Pahor et al., 2022).

Predictor variables

Before training, participants completed several cognitive assessments and a series of self-report questionnaires capturing demographic information including SES and video game playing experience, personality traits and other characteristics like growth mindset. All assessments and questionnaires were conducted on tablets and/or via Qualtrics

software. The descriptive information for all predictors is provided in supplementary materials Table S1.

Cognitive Assessments

WM updating. We employed an untrained variant of the N-back task to measure participants' updating skill, which served as a proxy of specific training task baseline measure, using either pictures of animals or vehicles (Pahor et al., 2022). All participants completed 1-back, 2-back, and 3-back levels (in that order). Each N-level consisted of 30 trials with nine targets. Each stimulus was presented 2.5 seconds with a 500 ms interval. Accuracy for each N-level was calculated as hits/(hits+misses+false alarms). We averaged z scores of 2-back and 3-back accuracy as the predictor variable.

WM capacity. To examine WM ability in different contexts, we used two WM tasks. Letter-number task (LN) was a verbal WM task similar to the letter-number sequencing subtest of the Wechsler Adult Intelligence Scale-III (Wechsler, 1997). Participants were required to remember and sort a mixed order of letters and numbers. For instance, the sequence 'H8T3K5' would be sorted into '358' and 'HKT'. Set size was the length of the sequence, and the task started from set size 2 and ranged to 15. Two trials were presented per set size. The task ended if both trials on the same set size were recalled incorrectly. The outcome variable was the highest set size achieved where at least one trial was correct. Corsi Blocks Forward (CF) was a visuospatial WM task inspired by traditional e-Corsi Block tests (cf. Ramani et al., 2020). In the task, participants saw a sequence of gophers appear in different holes; after that, participants were asked to tap on them in the correct order. Each gopher was displayed for 1.5 seconds and the interstimulus interval was 0.5 seconds. Set size ranged from 2 to 10, and the task began with the lowest set size. Similar to the letter-

number task, there were two trials per set size, and the game ended when both trials were recalled incorrectly. The outcome variable was the highest set size achieved. We used the average z score of both tasks (LN and CF) to reflect WM baseline capacity.

Inhibitory control. We tested inhibitory control (IC) as a predictor given extant research suggesting that IC is highly correlated with WM (Robert et al., 2009), however its role in WM skill learning is to date unknown. We used a countermanding task, which is a modified version of the Simon and spatial Stroop task, measuring the interference caused by spatial incongruent stimuli (Davidson et al., 2006; Diamond, 2013; Wright & Diamond, 2014). In this task, participants were presented with visual stimuli in the form of pictures of dogs and monkeys appearing on either the left or right side of the screen. They were instructed to tap on one of two green buttons in response to these stimuli (Ramani et al., 2020). For dogs, participants were required to tap on the button that was on the same side of the screen (congruent trials); however, when the participants saw a monkey, they were required to tap on the button that was on the opposite side (incongruent trials). The outcome measure was the reverse z score of the average response times spent on incongruent trials minus the time on congruent trials, thus a higher z score reflecting higher IC ability.

Fluid reasoning. We used the University of California Matrix Reasoning Task (UCMRT) to capture fluid reasoning skills (Pahor et al., 2019). Participants were shown a 3×3 matrix of stimuli with a missing element. They were required to select the missing part from eight answer alternatives to complete the pattern. There were 23 problems with a time limit of 10 minutes. The predictor variable was the proportion of correct answers.

Person-Related Factors - Self-Report Measures

Growth mindset. We used the Growth Mindset Scale (Dweck, 1999) to assess participants' beliefs about the malleability of intelligence. For our study, we replaced the word "intelligence" with "cognitive ability", e.g. "No matter who you are, you can significantly change your cognitive ability". In Study 1, participants responded on a 6-point scale from "Strongly disagree" to "Strongly agree". However, in Study 2 and 3, the "Strongly disagree" option was not displayed by mistake. Thus, we only included data from study 1 and calculated the sum of 5 items as the predictor variable.

Cognitive failures. To assess self-reported failures in memory, we selected 8 items from Broadbent et al.'s cognitive failure questionnaire (CFQ) (1982) as modified by McVay and Kane (2009). Participants were asked to report how frequently they encountered each scenario in the last twelve months (e.g. "Do you forget to give a message to somebody as you were requested to do?") on a Likert scale from 1 (Never) to 5 (Very often). Sum of all items was used as a predictor variable.

Grit and Ambition. Grit is described as a positive trait toward passion and perseverance. There were 8 statements to assess grit (Duckworth & Quinn, 2009), e.g. "I finish whatever I begin". Participants evaluated each statement with 5 points from "Very much like me" to "Not like me at all". Ambition is a concept showing the desire for success and captured by 5 items (Duckworth et al., 2007), like "I am a hard worker" using a 5-point scale as Grit. Since ambition is commonly related to grit ($r = .518$ in our dataset), we used a composite score averaging across the two scales.

Personality Traits. We used the 40-item Mini-Markers questionnaires (Saucier, 1994) to capture participants' big five personality traits (Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness). Each personality dimension was represented by 8

items consisting of adjectives. Participants were asked to rate how true the items are about themselves using a 5-point likert scale. Each trait was represented by its average score across corresponding items, serving as a predictor variable.

Environmental Factors and Experience

Socio-economic Status (SES). SES was inferred by both participants' self-reported subjective SES and their parents' education level. We used the MacArthur Scale of Subjective Social Status–Youth Version (Goodman et al., 2001) in which participants were shown two 10-step ladder pictures representing their community and the whole country. Then, they were asked to rate where their families would be on ladders. From the top of the ladder to the bottom were coded from 10 (the highest SES) to 1 (the lowest SES). The overall score of the two ladders reflected their self-reported SES level. In our dataset, the overall score ranged from 4 to 20 (average level = 11). Parents' education level (less than high school, high school diploma or equivalency, associate degree (junior college), Bachelor's degree in college, Master's degree, Doctoral degree, professional degree (JD, MD), and other specifics) was also collected and coded from 0 to 6. The sum of both parents' education coding scores was from 0 to 12 (average score = 4.75) in our dataset, indicating a wide range of SES. Both parents' education level along with self-reported SES level were used as predictors.

Video Game Background. We assessed participants' gaming experience using a video game questionnaire (VGQ) as used in previous work (Waris et al., 2019). In our survey, participants estimated how many hours they spent in 6 categories of video games (Action with shooters like Call of Duty, Action with adventure like Grand Theft Auto, Non-action role playing game like World of Warcraft, Real-time strategy like Starcraft, Turn-based

strategy like Civilization, Music games like Guitar Hero) per week in the past year. Their estimations were done using a 6-point scale that included Never, 0 to 1 hour, 1 to 3 hours, 3 to 5 hours, 5 to 10 hours, 10+ hours. Consequently, we converted the individual sum of video game hours within the populations of Study 2 and Study 3 into z-scores. We also calculated z-scores within Study 1. Finally, we merged the z-scores obtained from all three studies.

Health Status. A total of 3 self-rated items were used to assess participants' satisfaction with their health (e.g. How satisfied are you with your present physical health/psychological health/physical fitness?). Each item is assessed by a 5-point scale, from 1: Very dissatisfied to 5: Very satisfied. The raw scores served as 3 predictors.

Bilingualism. We created a dummy variable to indicate whether or not participants were bilingual (0 = monolingual, 1 = bilingual). 80% of our sample identified as bilingual (N=403).

Analytic plan

The data analysis comprised two primary steps: preprocessing and training the machine learning models. Preprocessing involved handling missing values and outliers separately in the training data and predictors. Following that, participants were grouped based on the training data, and the machine learning models were trained to predict their respective groups (refer to Figure 1.1 for the flowchart).

Training performance preprocessing

To identify outliers during training, we calculated each participant's average n-back level and standard deviation (SD) across training sessions. Since the distribution of SD and average performance were skewed (Skewness=8.28 and 7.77, respectively), outliers were

identified by a skewness-adjusted boxplot tool (Hubert & Veeken, 2008). Afterward, we excluded 20 outliers whose average performance or SD did not fall within the skewness-adjusted .25 to .75 quantile range (mean of outliers' average n-back level=16.25, SD=7.50). After deleting outliers, we applied piecewise linear models to fit each participants' training trajectory. We divided the training curve of each participant into two parts, each part was fitted with a linear curve, and the two straight lines were connected at a knot (Ørskov et al., 2021). We restricted the position of the knot not to appear in the first or last session. We were not able to fit 42 participants' training data with the piecewise linear model because their performance did not change across all sessions, i.e., they essentially did not show any learning. Further analysis excluded 11 of them since they had only been exposed to the 2-back task throughout the entire training period due to a specific algorithm setting. The remaining 31 participants who kept flat during training sessions (i.e. starfish group) were treated as a specific group discussed in supplementary materials. For each of the remaining 506 participants, we used a data-driven approach to put the knot at different time points (i.e. sessions) and determined the optimal knot location by the best model fit (largest r square). With the optimally fitted knots, the mean r square across all participants was .73. We also used linear function, quadratic function, exponential function, and logarithmic linear function to fit participants' training trajectories, however, the average model fit was not as good as the one using piecewise linear models (average r square = .47, .62, .44, and .52, respectively). After fitting training curves with piecewise linear models, we observed that the knot locations of the curves varied among participants (Mean knot = 5.13, Skewness = .18). Additionally, the slopes of the first part (slope1) also showed variation among participants (Mean slope 1 = .45, Skewness = 2.55). In contrast, the slopes

of the second part (slope2) were centered around 0 (Mean slope 2 = .00, Skewness = 5.43). The informative variation in the population seems to be concentrated in the knot and slope 1. For this reason, we used participant's knot location and only slope1 to represent training trajectory.

We decided to identify different learning patterns using these two indicators (i.e., knot location, slope1) instead of directly utilizing the continuous indicator values. This decision was made due to the reason that when we attempted to use slope1 as the target and trained regression models (excluding the no change group), the prediction performance yielded RMSEs above 0.35 (see supplementary Table S4). This outcome suggests that our predictors might not have sufficient sensitivity to accurately distinguish subtle differences in learning rates. Therefore, we employed clustering to identify learning patterns from both the participants' knot location and slope1, and treated participants who did not exhibit any changes during training as a single learning pattern.

Clustering

Clustering is an unconstrained method that does not make any assumption of the performance change curve and helps to determine the best grouping among participants. To detect different training patterns underlying participants' training trajectories, we used the K-means algorithm, which is commonly used in clustering (Likas et al., 2003). The K-means algorithm is intended to partition data into k clusters so that data points (participants in our case) within the same cluster are more similar, and points in different clusters are more dissimilar. Similarity between two participants was determined by the Euclidean distance between them, specifically, calculated by their knot location and slope1. K-means tries to minimize distances within a cluster and maximize the distance between

different clusters. We implemented K-means to divide participants into three clusters based on their knot location and slope1 across training (both standardized). To measure the goodness of the clustering result, the average silhouette score was computed. Silhouette scores are used to measure how dense and well-separated the clusters are. Silhouette scores range from -1 to 1, with higher scores indicating that the clusters are clearly distinguishable. The silhouette score of the three classes in our dataset was .47. We tested alternative clustering into two or four clusters (Silhouette score =.49 and .45, respectively) or more clusters (see supplementary Figure S2), and even though the silhouette score using two clusters was numerically higher than the three-cluster solution, we decided to work with the three clusters given that the two-cluster solution would likely fail to capture more subtle training patterns. Afterwards, participants were divided into three groups as determined through clustering, consisting of 45, 236, and 225 participants, respectively (see Figure 1.1). For the remainder of the paper, we refer to the participants in those three groups as "Unicorns", "Tortoises", and "Hares", which captures their training trajectories (Figure 1.2).

Predictors cleaning

Since many machine learning algorithms do not support missing values, we addressed missing values on predictor variables first. First, we checked the missing rate for each predictor. Growth mindset was excluded because it had about 61% missing values due to the aforementioned errors with administration during study 2 and study 3. Missing rates of other predictors ranged from 0 to 13.7%. Second, we looked at the missing rate for each participant. We conducted a Little's Missing Completely at Random (MCAR) test, which resulted in a non-significant value of .778. This suggests that the missing pattern in

our data was completely random. Nonetheless, the proportion of missing data is directly related to the quality of statistical inferences (Dong and Peng, 2013). To omit obvious loss of information, we decided to exclude participants who did not complete 50% of the survey with above nine missing values out of 18 predictor variables. The remaining missing values were imputed using Multiple Imputation by Chained Equations (MICE), a method in which each feature with missing values is modeled as a function of its neighboring features in a round-robin fashion (Zhang, 2016). Specifically, MICE was implemented using the multivariate imputer with the Bayesian Ridge regression model. The imputer followed an ascending imputation order and considered all features. The proportion of imputations (number of imputed cases/number of total cases) is provided in Table S1. Afterwards, we combined the cleaned predictor variables of the participants with their clustering results. Since we did data cleaning for training performance and predictors in parallel, we needed overlapping cleaned cases that do not contain outliers or excessive missing data in both intervention and predictor variables. In total, there were 472 cases left for classification.

Classification preprocessing

Classification preprocessing is the first step in the application of machine learning algorithms (Albon, 2018). The objective is to increase the signal-to-noise ratio and eliminate redundant or irrelevant information. Our preprocessing included Train-test split, Standardization, Oversampling, and Feature selection. Train-test split is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. We used a stratified shuffle split method randomly picking 20% participants as a test set having equally balanced classes. The remaining 80% of participants were used as a training set to feed classification models. All predictors in

both train and test sets were standardized separately to keep them in the same range and treated the test set as a completely new dataset. Specifically, the train set was standardized within the set, while the test set z score was calculated based on the mean and standard deviation from the train set: (value from test set - corresponding mean from train set)/(corresponding standard deviation from train set). By doing so, we utilize information from the train set to better scale the test set due to the limited sample size of the test set. We used oversampling and feature selection as strategies given that we were dealing with a relatively small data set for machine learning field. Oversampling was used to prevent the reduction of the model's predictive power caused by an unbalanced data set. In this case, we used the SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous). SMOTENC creates new synthesizing predictors' data for the minority class, based on those that already exist (Chawla et al., 2002). It can handle dataset containing numerical and categorical features. We used oversampling to increase the sample size of the minority group to be the same as the majority group in the train set. Feature selection allows us to remove unneeded features that do not contribute to prediction and to avoid overfitting (Guyon & Elisseeff, 2003). Specifically, we used a robust feature selection method on the train set: L1 regularization or so called LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996). With the LASSO method, the regression coefficients are reduced to zero, which regularizes the model parameters and avoids overfitting. Features with non-zero coefficients will be kept in training models. To choose the best regularization parameter C (inverse of regularization strength), we explored values between 0 and 1, with the best C set to be 0.1. To determine the effectiveness of

feature selection, we compared the model performance with all features included. The results can be found in the supplemental materials, Table S5.

Binary tree model

The clustering procedure assigned each participant a label, which was the target we needed to predict with the machine learning model. This known target prediction process required the use of supervised classification learning models. Specifically, we used a hierarchical binary tree model to divide multiclass classification into two binary classification models (see Figure 1.1). We chose a binary tree model over a multiclass model due to the varying performance observed between the labels. While we initially attempted a multiclass model, it did not yield satisfactory results, likely due to the significant differences in distinction between the unicorn and non-unicorn group, as opposed to the distinction between the tortoise and hare group. Hence, we decided to employ two separate algorithms to effectively differentiate these label groups: The target of model 1 was to detect unicorn group, whereas the target of model 2 was to distinguish between the tortoise and hare group. The two models were trained separately with the same classification preprocessing approach. For each model, after feature selection, we first applied a variety of algorithms with default parameters to train the data in order to avoid personal bias in choosing a classification model (Géron, 2019). This was achieved via a model selection function in the scikit-learn library (version 1.1.2) in Python (Pedregosa et al., 2011), which evaluates several algorithms' performance (e.g., Logistic Regression, Decision Tree, etc, cf., supplementary materials; Table S6). The evaluation approaches were discussed in the next paragraph. Model 1 and model 2 were selected based on different criteria. For Model 1, our primary focus was on accurately detecting unicorns and avoiding

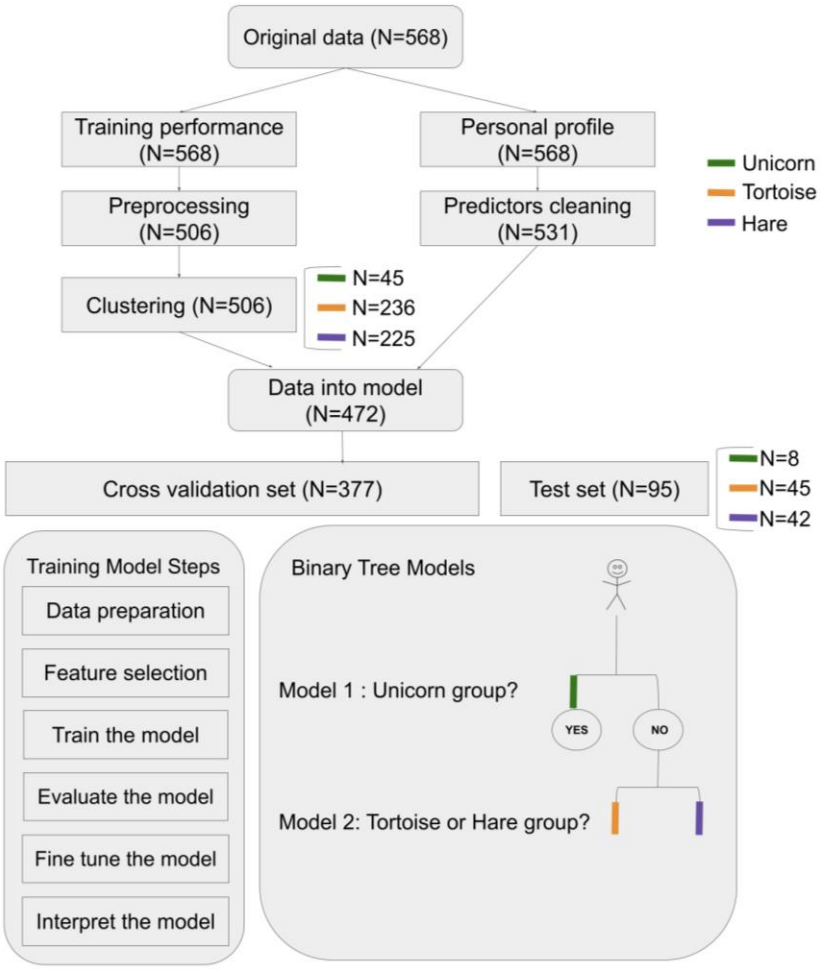
the omission of positive instances. On the other hand, Model 2 aimed to differentiate between hares and tortoises with utmost accuracy. We found the Radial Basis Function (RBF) Kernel Support Vector Classifier (model 1) and Random Forest algorithm (model 2) to be the most promising models. RBF kernel support vector classifier is a method that is responsible for finding a non-linear decision boundary to separate different classes and maximize the distances between cases at boundaries (Schölkopf et al., 2004). It can handle nonlinear data with high dimensions. Random Forest is an ensemble model based on a number of Decision Tree classifiers (DT). Each DT will learn from data and make a flow of if-then-else rules to predict new instance labels. RF then averages all trees' decisions and makes a final prediction. Random Forest is also a commonly used approach in precision education using machine learning (Luan & Tsai, 2021). Random Forest has advantages in dealing with unbalanced data and avoiding overfitting by restricting how many if-then-else rules in each DT. Random Forest has no assumption of data structure that can capture nonlinear relationships and allow interactions among predictors (Denisko and Hoffman, 2018).

After model selection, we applied an exhaustive search over specified parameter values for two models and selected the optimal parameters according to the model performance. When evaluating the performance of the model, we used the cross-validation method, specifically the 10-fold method. Ten-fold cross-validation is an estimate of the generalization performance of a model trained on 9/10 samples (Berrar, 2018). Unlike the traditional division of data into a training and testing set, 10-fold cross-validation treats each fold as a testing set, and the remaining participants as the training set. In that regard, we trained and validated a model 10 times and looked at their average performance. In

addition, we evaluated model performance with widely acceptable performance metrics including A) a confusion matrix, which indicates the combination of true labels and predicted labels; B) The mean accuracy across all 10 folds, evaluated using the f1_weighted scoring metric that considers the performance across all classes; C) F1 score as a weighted harmonic mean of precision and recall calculated based on confusion matrix (for the formula see supplementary materials), with scores ranging between 0 and 1 (scores closer to 1 indicate better classifiers). Due to the hierarchical binary tree model we used here, the chance level for prediction is .50 for each binary model within it. When we apply this binary tree as a whole, the chance level for predicting membership in any of the three groups is approximately .33. It is worth noting that among our training data, some of them were created by the oversampling method (~50% for model 1). This part of the data was just to balance each class to help train better models. Therefore, when validating the model, we eliminated those synthetic data so that all the model performance indicators only involve the original dataset.

Figure 1.1

General Flowchart of Analytic Plan



Note. Three colors represent different groups identified by clustering.

Model interpretation

After training the machine learning model, we applied a method called SHAP (SHapley Additive exPlanations) to determine the extent to which each feature contributes to making predictions in final models (Lundberg and Lee, 2017). This method can demonstrate feature importance, both globally and locally, to supervised learning of labeled data. The SHAP are based on concepts from cooperative game theory, which assigns

each feature an importance value based on whether the feature is present or not during SHAP estimation (Lundberg and Lee, 2017; Lundberg, Erion & Lee, 2018). Specifically, for each case, the final model provided the predicted probability of belonging to a specific label. The difference between this probability and the baseline probability (the label's proportion in the population) was quantified by SHAP values, indicating the predictors' contribution. A positive SHAP value indicates a positive contribution, meaning it tends to predict the label, while a negative value indicates a tendency to not predict the label. The important predictors are those with higher average absolute SHAP values across all cases. Since different explanation formulas are suitable for different types of models, we utilized a kernel-based explainer for the RBF Kernel Support Vector Classifier (model 1), and a tree-based explainer for the Random Forest Classifier (model 2).

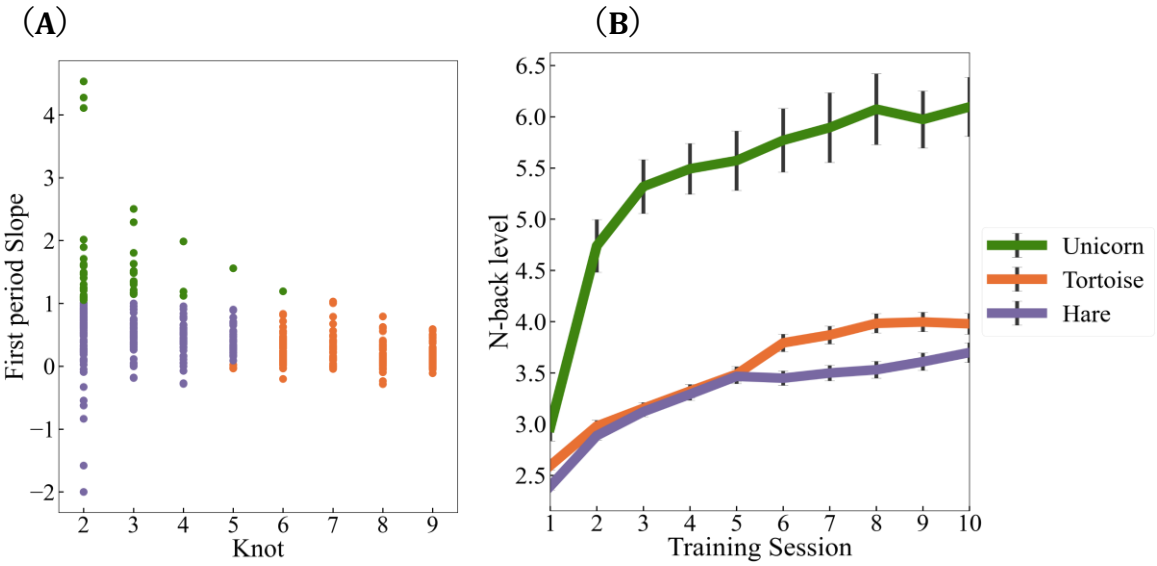
1.3 Results

Learning pattern

As shown in Figure 1.2, the three colored curves represent the average training trajectories of the three learning patterns as determined by the K-means method. Unicorns outperform both other groups, showing a much higher performance from the beginning and extending their high performance throughout training; Tortoises maintain a regular rate of daily progress compared to hares, which reach an early asymptote. Mean training curves for each group are shown in the supplementary materials (Figure S3).

Figure 1.2

Learning Patterns Identified by Clustering



Note. A) Scatter plot of participant's optimal knot location (x axis) and initial training gain (slope1; y axis) as a function of the three learning clusters. B) Average learning trajectories of the three clusters across the 10 training sessions. Error bars represent ±1 standard error of the mean.

The results of subsequent ANOVA analyses revealed that participants started at different n-back levels during session 1 with significant differences emerging between groups at the outset ($F_{2,503} = 16.98$, $p < .001$). Bonferroni post hoc tests confirmed that the unicorn group (Mean WM span = 2.97) significantly outperformed the tortoise (Mean WM span = 2.59) with $p < .001$, *Cohen's d = 0.53* and hare groups (Mean WM span = 2.39) with $p < .001$, *Cohen's d* = 0.94. The unicorn group also showed a high growth rate (Mean slope 1 = 1.63) over the first three sessions (Mean knot location = 2.62), and kept its high performance (Mean slope 2 = 0.07). The tortoise and hare groups showed slower growth

rates (Mean slope 1 = .22 and .45, respectively), and tortoises progressed significantly more slowly than hares (*Cohen's d = 0.77*). The tortoise group, however, continued to improve for a longer period of time than the hare group (Mean knot = 7.42 versus 3.23, *Cohen's d = 3.52*). Finally, the tortoise group performed slightly better than the hare group at the end of training (performance at session 10: 3.98 versus 3.70, *Cohen's d = 0.19, p = .047*).

Model performance

With this model, we applied the feature selection method described above to determine the optimal subset of predictors. WM, Fluid reasoning, Inhibitory control, Grit and Ambition, Consciousness, Openness, Extraversion, Neuroticism, SES, Video game background, Health status, and Bilingualism were included in the final binary model 1; Inhibitory control, Cognitive failure, Openness, Neuroticism, Extraversion, and Video game background were included in the final binary model 2.

As shown in Table 1.1, two models performed well on the cross validation set. Model 1 (RBF Support Vector Classifier), when identifying unicorn participants, achieved an average accuracy of .90 across 10 folds validation sets (*SD* = .03), with an F1 score of .46. Model 2 (Random Forest Classifier) showed less power in predicting whether participants were from the tortoise or hare groups. The average accuracy of .64 across 10 folds (*SD* = .05), with an F1 score of .65.

Table 1.1

Confusion Matrix of Binary Tree Model Performance on Cross Validation Set

		Predicted class	
	Model1 (N=377)	Unicorn	Non-Unicorn
True class	Unicorn	**28 (.88)**	4 (.12)
	Non-Unicorn	62 (.24)	**283 (.82)**
	Model2 (N=345)	Tortoise	Hare
True class	Tortoise	**114 (.64)**	63 (.37)
	Hare	62 (.36)	**106 (.63)**

For the purpose of evaluating the generalizability of our binary tree model, we determined its accuracy when applied to a separate set of holdout test data. As a first step, all data from the cross-validation training set was fitted into the final binary tree model with optimal hyperparameters, and then we predicted whether or not test set participants belonged to the unicorn group. Participants predicted as non-unicorns were assigned to the second binary model. Using the second binary model, we were able to predict whether those non-unicorns belonged to the tortoise or hare group. Based on the results presented in Table 1.2, Model 1 was able to detect 5 out of 8 participants in the unicorn group (model accuracy = .74 over chance level .50). Model 2 correctly identified 57% of the non-unicorn group as belonging to the tortoise or hare groups (model accuracy = .59 over chance level .50). Overall, in the test set, after applying the final binary tree model, about 51% participants' learning patterns were correctly predicted over chance level (33%).

Table 1.2

Confusion Matrix of Binary Tree Model Performance on Test Set

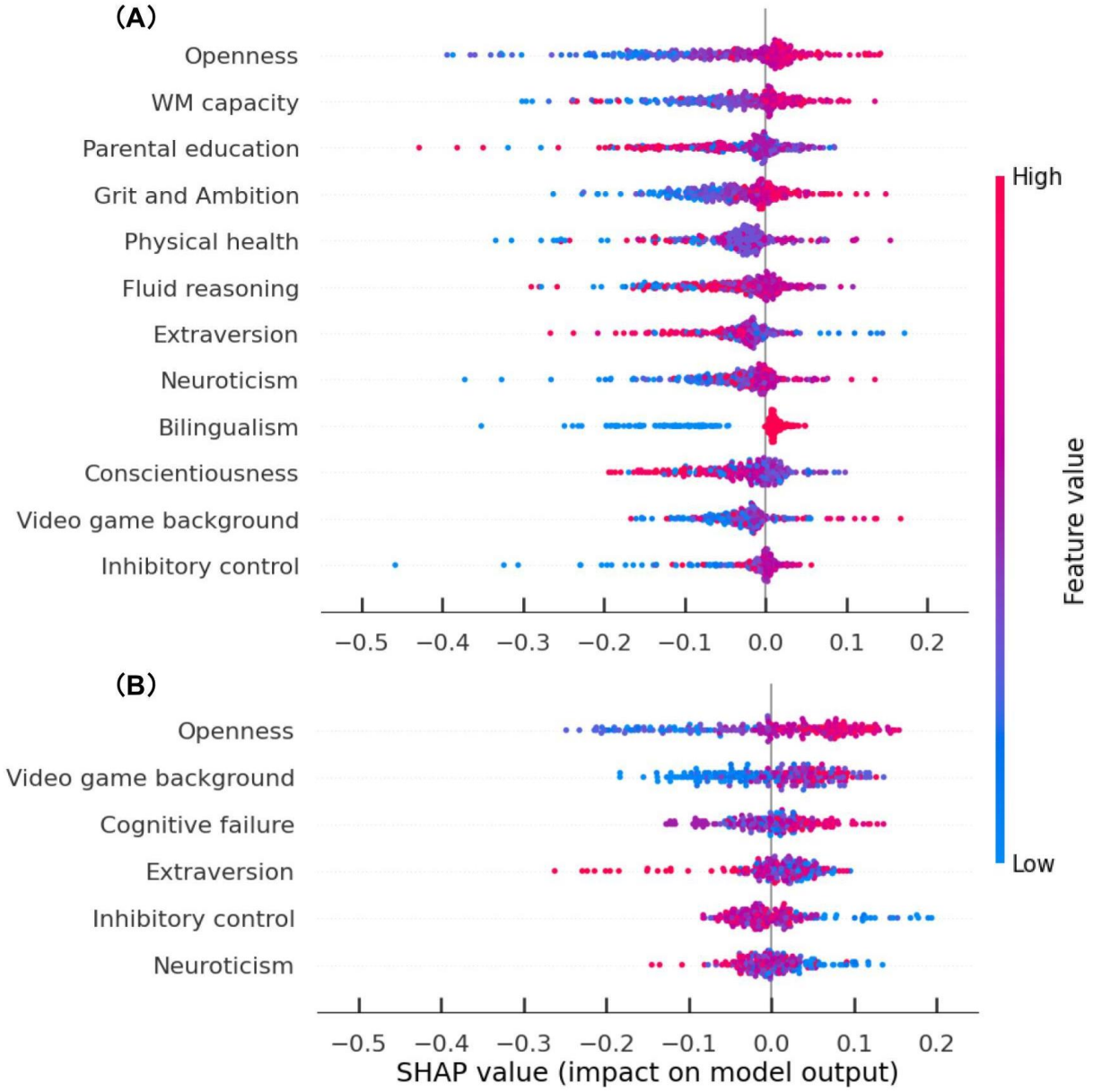
		Predicted class	
	Model1 (N=95)	Unicorn	Non-Unicorn
True class	Unicorn	**5 (.63)**	3 (.37)
	Non-Unicorn	14 (.16)	**73 (.84)**
	Model 2 (N=76)	Tortoise	Hare
True class	Tortoise	**21 (.57)**	16 (.43)
	Hare	14 (.39)	**22 (.61)**
	Unicorn	2 (.67)	1 (.33)

Predictor importance

To interpret our final binary models, we calculated the mean SHAP value for each predictor to represent its feature impact (See Figure 1.3). Considering that the cross-validation set contains more data points than the test set, we examined the SHAP values of the cross-validation set. In addition, to focus on how the model successfully made predictions, the following visualizations only included those instances in the cross-validation set that were correctly predicted by our final models. A positive value indicates a positive contribution, that is, an increase of the probability of this person belonging to a particular class, while negative values indicate a decrease of the probability of belonging to that class.

Figure 1.3

Predictor Importance by Calculating the SHAP Value



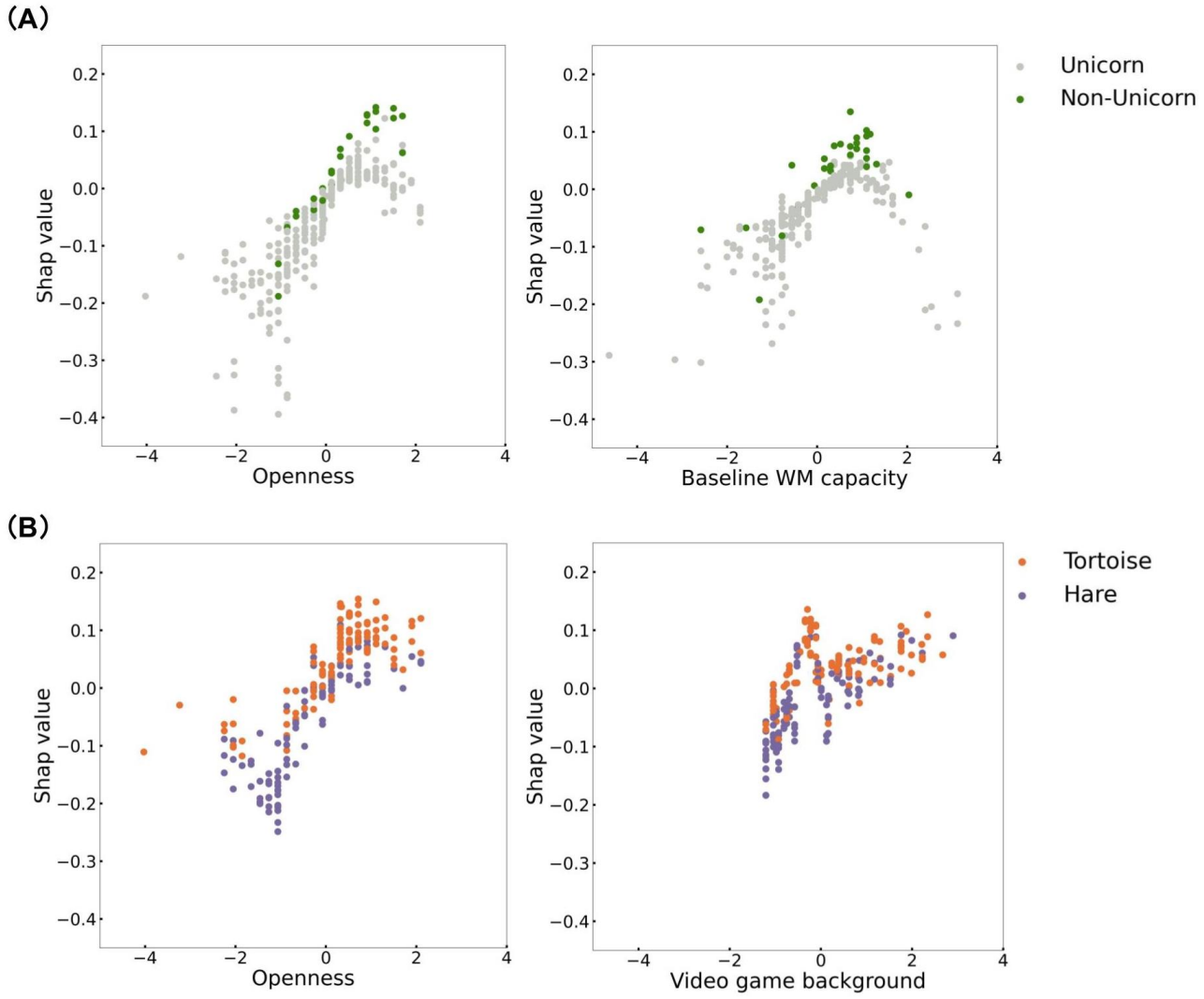
Note. The x-axis represents the SHAP value; The y-axis represents the predictors. Each dot represents each participant who is correctly predicted by our model (including both hitting a particular class and correctly rejecting the class), the color of the dot represents a certain predictor's value (red: higher value, blue: lower value). **A.** Each predictor's SHAP value in

predicting the unicorn group (model 1) ranking by predictor importance. **B.** Each predictor's SHAP value in predicting the tortoise group (model 2) ranking by predictor importance.

Figure 1.4A indicates the absolute mean of the SHAP values in predicting the unicorn group. Overall, openness and baseline WM capacity were the two most important factors contributing to the prediction. Specifically, participants who score higher in openness and with better baseline WM capacity were more likely to be classified as the unicorn group. Interestingly, openness and video game background emerged as the top two factors when predicting tortoise or hare groups (Figure 1.4B). Participants with a higher value of openness and more experience with video games were predicted to belong to the tortoise group. Figure 1.4 provides the scatter plots illustrating the relationship between the most important features values and their SHAP values.

Figure 1.4

Scatter Plots of Predictor Values (x axis) and SHAP Values (y axis)



Note. All predictor values were transformed to z scores. The top two important predictors in predicting unicorns (green) versus non-unicorns (gray) (Figure 1.4A) and tortoises (orange) versus hares (purple) (Figure 1.4B) are shown.

1.4 Discussion

Our study used machine learning models to examine whether and how a combination of individual difference variables could predict inter-individual variations in learning

trajectories in the context of WM training. We used a clustering approach to identify three different participant groups ("Unicorns", "Tortoises", and "Hares") relying on their early learning slopes and turning points during WM skill learning. The unicorns demonstrated high initial performance and quick learning, while the tortoises and hares displayed lower performance initially, with gradual improvements over time. Tortoises and hares differed in that tortoises continued to improve over longer periods, ending up with a relatively higher performance, whereas hares' performance leveled off earlier. We then developed a binary tree model to predict which learning patterns participants belonged to. Among our 18 individual difference predictors, openness, as well as baseline WM capacity were the highest contributing factors in the first binary model (RBF support vector classifier) when predicting whether or not participants were unicorns. When we generalized the first model to new samples, our predictive accuracy was .74. For the Non-Unicorn group, we applied a second binary model (random forest classifier) to predict whether they were hares or tortoises. In this second model, we found that openness and video game background were the most important factors, and the predictive power for the new samples was .59.

The three different learning trajectories that emerged in our analyses broadly support Ackerman's theory (2007) in demonstrating that differences between individuals become more pronounced with training. Additionally, the observed differences among the three patterns strongly support the magnification account (especially when comparing the unicorn group and non-unicorn group), with individuals initially demonstrating higher performance showing the most pronounced training gains, while those with lower initial performance exhibiting markedly less improvement (Lövdén et al., 2012; Rebok et al., 2007). This finding is consistent with a recent study by Ørskov et al. (2021) that used a similar

intervention (dual-n-back training) in an adolescent population. Several studies, including Ørskov et al., argue that strategy use might explain this magnification effect (Fellman et al., 2020; Könen & Karbach, 2015). According to those authors, high performers might be better at coming up with effective strategies for specific tasks, which leads to faster learning and better overall performance. An alternative account suggests that higher initial cognitive abilities may also determine the later stages of skill learning (Taatgen, 2001). That is, high performers are better at learning in general, which translates to steeper training slopes, which we observe in our data. Laine et al. (2018) argued that those who are more successful at WM training are likely to use training task-related strategies such as updating and grouping, thus, participants need to heavily rely on their task-related WM resources, especially at higher N levels. As such, if the magnification effect was driven by strategy use, we should have found that participants' baseline WM updating as measured by untrained n-back tasks to be among the highest contributing factors in predicting whether participants were high performers. Instead, our results showed that more general WM capacity (measured by several WM tasks) carried the most weight in the prediction, indicating that more general WM abilities seem to contribute more to learning than task specific performance and/or strategies, which is more in line with Ackerman's theory (Ackerman, 1992). In addition to testing the prediction power of specific task performance, we also tested the prediction power of training-related processes based only on the initial performance within the training tasks themselves. However, it turned out that initial training performance did not have a good predictive performance (see Table S8), indicating that the more general factors we collected before training were more powerful predictors than the task-specific performance at the beginning of training, again, supporting the notion that any

learnt strategies extend beyond the trained task, which is important because this might ultimately explain transfer (e.g. Pahor et al., 2022).

Consistent with our hypothesis, the training trajectories of participants were found to be best predicted by a combination of individual difference variables, including personality characteristics. Specifically, we observed a significant positive correlation between feature value and SHAP value, indicating that individuals with higher levels of openness were more likely to be predicted as unicorns. Openness is characterized by traits such as open-mindedness, creativity, and insightfulness. Individuals high in openness tend to embrace variety and diversity, as well as to be curious about their surroundings, and adopt creative approaches when faced with changing circumstances (McCrae & Sutin, 2009; Sutin, 2017). Such adaptive traits have been found to be positively correlated with deep learning instead of surface learning (Chamorro-Premuzic & Furnham, 2009). Consistent with this perspective, a recent study of a one-year long cognitive intervention in older adults demonstrates a greater improvement in learning among individuals with more adaptive personality traits (Kekäläinen et al., 2023). Interestingly, openness predicts the tortoise in a positive linear manner as well. It may indicate that open-mindedness and creativity are crucial elements for completing a relatively extensive training program with adaptive learning challenges. This might manifest in trying out different strategies to figure out the optimal one, and/or being more open to the idea that cognitive training might be beneficial, which might lead to more persistence. In addition, we noticed a trend indicating that participants with more video game experience were classified as "tortoises" who persisted longer throughout the intervention as compared to “hares” (Figure 1.4B). This aligns with previous research highlighting the cognitive benefits of video gaming (Steyvers & Schafer,

2020; Waris et al., 2019; Zhang et al., 2021), in particular, the 'learning to learn' framework (Bavelier et al., 2012). The repetitive nature of video games may also cultivate patience for cognitive tasks, enabling individuals to maintain their performance over time.

When using SHAP values to explain how the different factors predict learning patterns, we found that the impact of factors on learning trajectories is non-linear in most cases. As shown in Figure 1.4A, there is an inverted "U" shaped relationship between baseline WM capacity and its contribution to predicting unicorns. Importantly, this pattern confirms that individual differences can have non-linear effects on training performance, and machine learning is an effective tool to uncover these nonlinear relationships. Most participants in the unicorn group demonstrated higher baseline WM capacity compared to the average. However, it should be noted that individuals with the highest WM capacity were not necessarily classified as part of the unicorn group. This further demonstrates that a participant's training performance is determined by a combination of variables. Although we provide a ranking of predictor importance in Figure 1.3, this does not necessarily imply that the most significant factors explain the full variance. It is important to note though that even in our relatively large sample for psychology field, the unicorn group was the smallest in terms of sample size (n=45), and thus, typical intervention studies that have fewer participants might not include enough of such super learners or they might just see one single participant like this and might exclude them as an outlier. The use of machine learning enables us to detect and predict subgroups, which is a significant advantage over traditional regression analysis. At the same time, our findings highlight the importance of looking at subgroups when determining their contributing factors to learning outcomes, which requires large-enough sample sizes.

Overall, our study took a novel approach to exploring individual differences in WM training by using machine learning, which allowed us to go beyond explaining the linear relationship between one or few separate individual-difference variables and learning outcomes. At the same time, we acknowledge several limitations. First, despite the fact that our sample size was larger than that of most cognitive training studies (Bogg, & Lasecki, 2015), it was still small as compared to other studies that implement machine learning models (Youyou et al., 2015). In order to minimize the risk of overfitting, we used oversampling, feature selection strategies, and most importantly, a separate test set. Despite these strategies, our results might not generalize to the entire population. Since machine learning models using psychological experimental data are still rare, this study served first and foremost as a preliminary exploration. Second, our study used indicators from piecewise linear regression to describe training trajectories (Ørskov et al., 2021). Learning patterns that rely on other indicators might result in a different prediction model. Indeed, completely data-driven approaches are highly dependent on the type of data provided, and choosing the most appropriate learning pattern can be challenging. Third, due to time constraints, certain predictors, such as physical health, were measured using one item. This limited assessment may not fully capture the comprehensive representation of the predictor. Lastly, our combination of individual difference variables did not have a high level of predictive power. There are likely other factors that were not captured by our assessments, or there is unmodelable uncertainty that creates variation in the data.

Despite these limitations, our observation that baseline cognitive abilities are effective predictors of learning patterns is consistent with the existing literature and theory (e.g. Ackerman, 1992; Karbach et al., 2017). Furthermore, our study illustrates that different

combinations of individual difference variables can be used to distinguish between different types of learning patterns. This implies that the relationship between individual differences and WM training gains is complex, and that the most important factors may vary across different types of individuals. This may also explain why the results of previous studies have been inconsistent and further highlights the need to take individual differences into account, which requires large sample sizes. By identifying these predictors, we can potentially identify sub-populations that are more amenable to targeted interventions. Notably, we also observed a few participants who did not show any performance changes during training, which has often been overlooked in previous studies. Although the sample size was limited, we found that those participants' fluid reasoning performance was either extremely high or low (see Figure S4). Participants with extremely low reasoning performance may be less well suited for this type of cognitive training, while those with extremely high reasoning performance may find the training boring, resulting in a lack of engagement. However, it is worth noting that any task design (and gamification) will be inherently be more welcoming to some participants and less welcoming to others and so future research is needed to better address aspects of task design may be needed to reach different populations rather than concluding that some populations are less well suited for particular tasks. On the basis of this study, future research can develop more advanced models for predicting individual training trajectories or transfer effects over large datasets, which can ultimately inform the design of more effective and productive training programs. Such training programs could be then used to identify at-risk students according to individual differences and in turn, provide refined and personalized interventions, which is the purpose of precision education (Luan & Tsai, 2021).

Data availability

The data that support the findings of this study are openly available in OSF at https://osf.io/yuza3/.

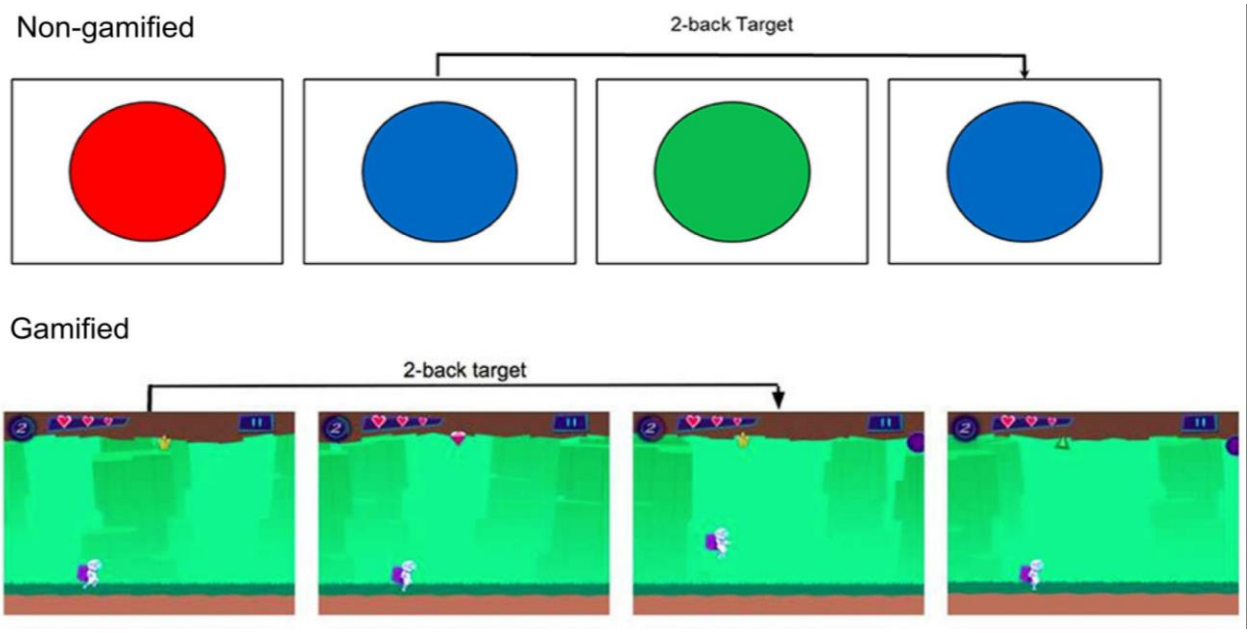
Acknowledgement

This work was supported by the National Institute of Mental Health (Grant No. 1R01MH111742; A.R.S. and S.M.J.) and by the National Institute on Aging (Grant No. 1K02AG054665; S.M.J.).

Supplementary Materials of Chapter 1

Figure S1

Non-gamified and Gamified Variants of the N-back Training Tasks



Note. The training paradigm was similar as described in Sandeep et al., 2020

Table S1

Descriptive Information for All Predictors, Along with Correlation Coefficients

Predictor variables	N	IR	Mean	SD	Min	Max	Kurt.	Skew.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Cognitive assessments																									
1. WM updating	565	0.2	-0.04	0.79	-2.23	2.17	-0.27	-0.09																	
2. WM capacity	566	0.2	-0.04	0.75	-3.62	2.37	1.96	-0.21	.21**																
3. Inhibitory control	567	0	0.00	0.92	-6.98	2.87	15.36	-2.35	.02	.044															
4. Fluid reasoning	567	0.2	.61	.28	0	1	-0.25	-0.54	.28**	.22**	.02														
Person-Related factors- Self-report measures																									
5. Growth mindset	220	/	20.02	4.54	6	30	-0.02	-0.38	.08	.08	.06	-.03													
6. Cognitive failure	527	0	24.19	5.42	8	40	-0.02	0.12	.10*	.04	-.01	.10*	-.01												
7. Grit and Ambition	528	0	3.60	0.57	1.91	5	-0.29	-0.21	-.05	.00	-.13*	-.18**	.24**	-.35**											
8. Neuroticism	527	0	3.37	0.70	1.5	4.88	-0.57	-0.08	.00	.09*	.03	-.08	.12	-.31**	.29**										
9. Extraversion	527	0	3.10	0.77	1.38	5	-0.60	-0.10	-.04	.05	-.01	-.09*	.11	-.15**	.38**	.10*									
10. Openness	527	0	3.69	0.62	1.13	5	0.28	-0.37	.07	-.01	-.01	-.05	.09	.03	.33**	-.02	.35**								
11. Agreeableness	527	0	4.05	0.59	1.63	5	0.54	-0.76	.06	.00	.04	-.03	.28**	-.16**	.33**	.28**	.29**	.29**							
12. Consciousness	527	0	3.72	0.65	1.75	5	-0.13	-0.46	.00	-.02	-.04	-.10*	.16*	-.34**	.56**	.22**	.20**	.24**	.31**						
Environmental Factors and Experience																									
13. Parents' education	521	0.6	6.04	2.90	1	14	-0.91	0.21	.11*	.03	-.02	.09*	-.07	-.03	-.07	-.06	.03	.10*	-.02	-.10*					
14. Self-reported SES	529	0.2	11.19	2.87	4	20	0.17	0.04	-.10*	-.09*	.06	-.02	-.06	.13*	-.24**	-.15**	-.24**	-.19**	-.20**	-.16**	-.24**				
15. Video game.	526	0.2	0.01	0.99	-1.16	4.00	0.06	0.87	.11*	.10*	.00	.19**	-.01	.06	-.13*	-.00	-.06	.00	-.10	-.16**	.02	-.06			
16. Physical health	490	9.2	3.25	0.74	1	5	0.54	0.37	.07	-.01	-.10	-.09*	.05	-.08	.21**	.07	.21**	.16**	.08	.11*	.19**	-.28**	-.09*		
17. Psychological health	490	9.2	3.46	0.89	1	5	-0.43	-0.11	.04	.05	-.01	-.10*	.02	-.21**	.23**	.39**	.30**	.04	.17	.13*	.02	-.26**	.02	.26**	
18. Physical fitness	490	9.2	3.06	0.91	1	5	-0.31	0.04	.06	.05	-.09	-.03	.10	-.13*	.33**	.13*	.30**	.19**	.15	.21**	.12*	-.29**	-.11*	.65**	.28**

Note. IR = Imputation Rate (%); Kurt. = Kurtosis; Skew. = Skewness; Video game. = Video game background

Growth mindset was excluded because it had about 61% missing values.

* = p < .05, ** = p < .01

Three predictors utilized a composite score derived from different measures. The correlation between these measures and their respective composite scores is reported as follows: 1) WM Updating: The accuracy of the 2-back task showed a significant correlation with the accuracy of the 3-back task (r = .314**). The composite z-score exhibited a strong correlation with 2-back accuracy (r = .826**) and 3-back accuracy (r = .795**). 2) WM Capacity: WM capacity was measured using the Letter-Number task and the Corsi Block Forward task, which exhibited a correlation of .131**. The composite z-score demonstrated a high correlation with Letter-Number span (r = .745**) and Corsi Block Forward span (r

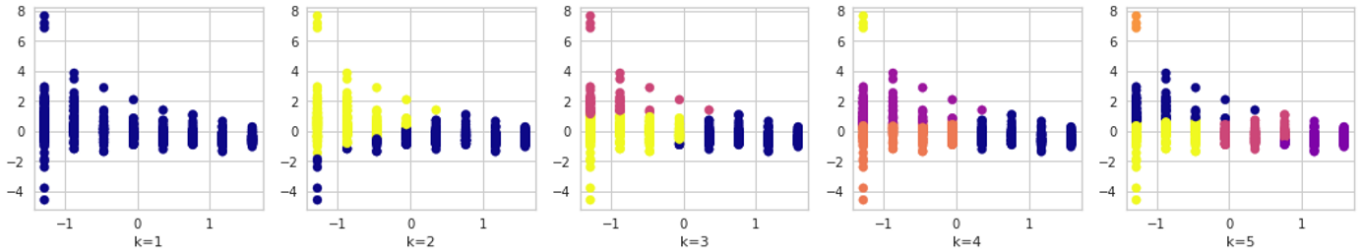
= .759**). 3) Grit and Ambition: Ambition was commonly associated with grit in our dataset, displaying a correlation of .518. The composite score of grit and ambition showed a strong correlation with ambition (r = .897**) and grit (r = .843**).

Training performance

Figure S2

Clustering Results with Different K Numbers

(A)



(B)

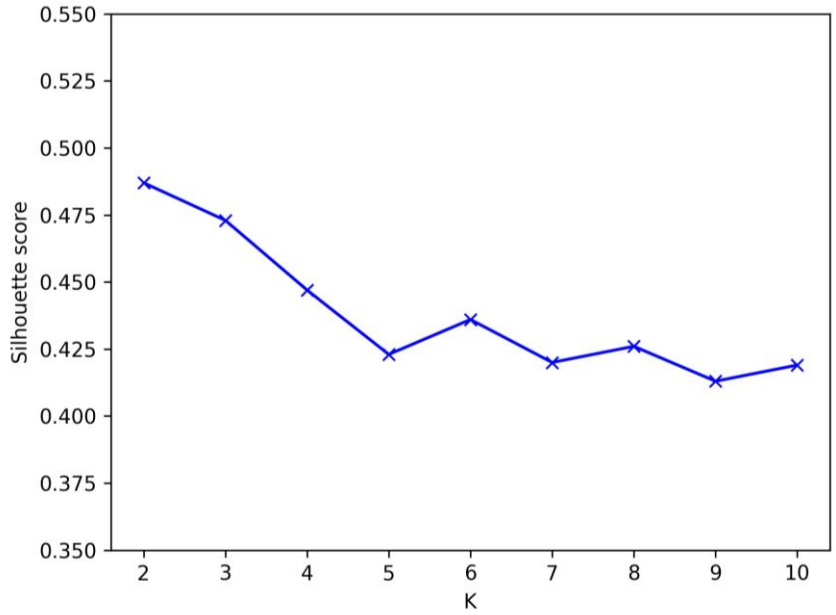
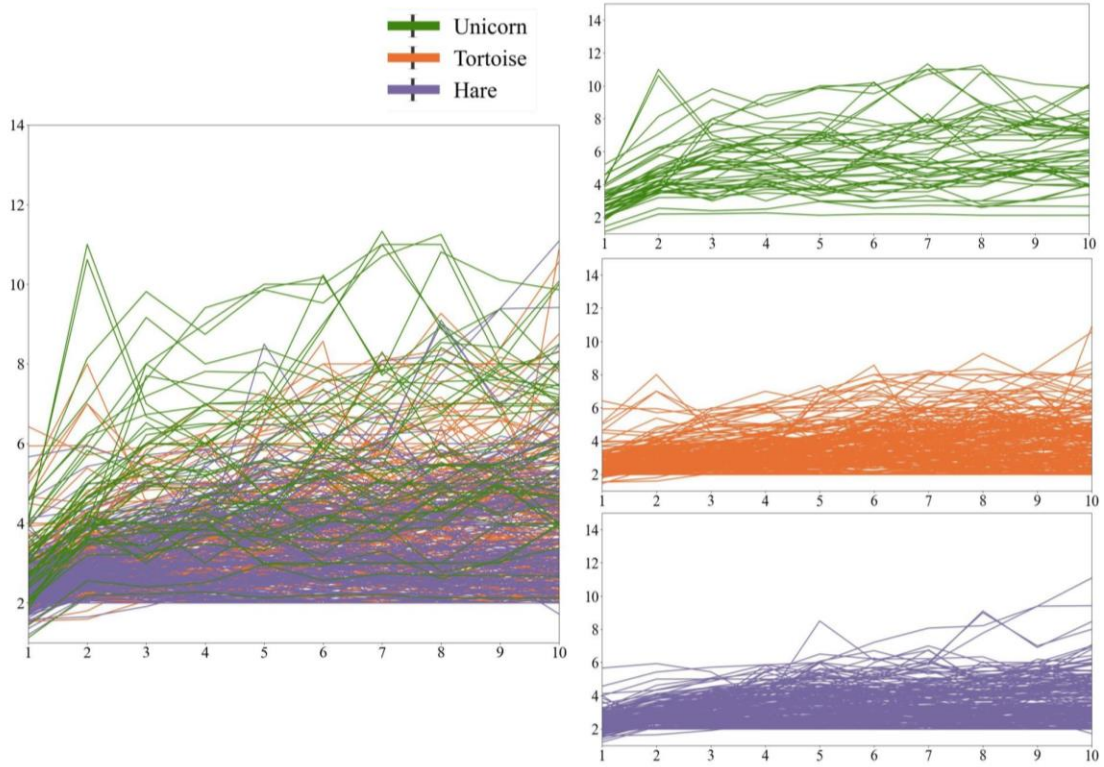


Figure S3

Training Trajectories by Cluster



Participants that showed no change during training

In our sample, there were 31 participants who did not show any performance change over the course of the training. We referred to this group as the "starfish group" given their lack of movement. Among them, 27 participants consistently performed at the 2-back level in each training session, while the remaining 4 participants consistently performed at the 3-back level. In order to investigate whether there are predictors that can distinguish between participants in the starfish group and those in the non-starfish group, we trained a binary model following the same model training steps as described in the main text. The results revealed that the RBF Support Vector Model performed the best on the cross-

validation set (Table S2), with an average accuracy of .94 (*SD* = .02). However, the test set showed poor accuracy (.17) since there were only 6 starfish participants in the test set. Still, we calculated the SHAP values in the cross-validation set to get some insight of which predictors might be able to distinguish between the starfish and non-starfish group. The results showed that fluid reasoning was the dominant predictor in the model (mean |SHAP value| = .21), followed by video game background as the second important predictor (mean |SHAP value| = .05). Interestingly, most of the participants who showed no performance change during training showed either extremely high or low performance on the fluid reasoning task (cf. Figure S4, and Table S3), suggesting that the reason for the lack of learning might be either related to motivation (e.g., those with high scores in fluid intelligence) or related to ability and/or motivation (those with low scores in fluid intelligence).

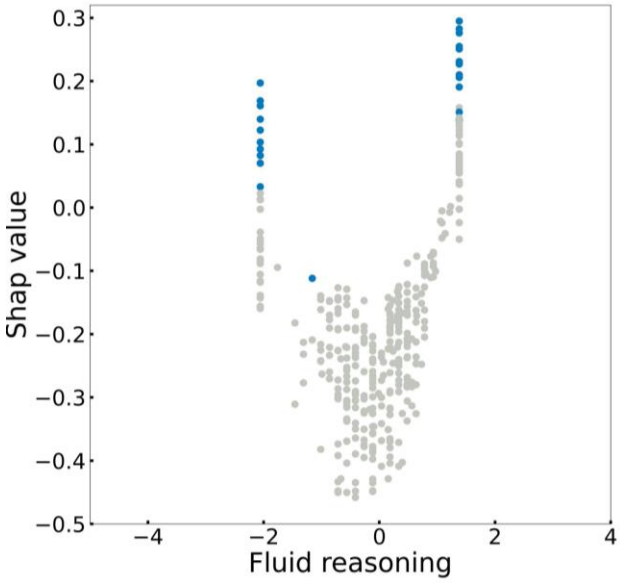
Table S2

Confusion Matrix: Model Performance in Detecting Participants with No Change during Training (Cross-Validation Set)

		Predicted class	
	Model (N=401)	Starfish	Non-Starfish
True class	Starfish	**24 (1.00)**	0 (.00)
	Non-Starfish	31 (.08)	**346 (.92)**

Figure S4

Scatter Plot of Fluid Reasoning and its SHAP Values



Note. x-axis: Fluid reasoning standardized score; y-axis: SHAP values. Blue dots: Starfish group who did not show any performance change during training; Gray dots: non-starfish group.

Table S3

Frequency Table of Starfish Group's Fluid Reasoning

	N	Fluid Reasoning (Accuracy)
Consistent at 2-back	12	0
	1	.217
	2	.261
	1	.619
	11	1
Consistent at 3-back	1	0
	3	1

Exploration of regression models

Since we had a continuous distribution of participants' learning slope rather than distinct clusters, we also trained regression models to directly predict their learning slope. The results of different models showed similar performance, see Table S4. The results demonstrated that these regression models exhibited a limited predictive power.

Table S4

Regression Model Performance (test set)

Model	MAE	MSE	RMSE
Linear Regression	.292	.133	.365
RBF Support Vector Model	.272	.128	.358
Decision Tree	.285	.128	.359
Random Forest	.291	.130	.360
KNeighbors Regressor	.348	.202	.450
Multilayer Perceptron Regressor	.322	.182	.426

Note. MAE: Mean Absolute Error; MSE: Mean Squared Error; RMSE: Root Mean Squared Error.

Model performance with all features

To assess the necessity of using feature selection, we trained a binary tree model using all predictors, excluding the training program information. The results are shown in Table S5. The performance of Model 1 was slightly better than the performance with feature selection, while Model 2 showed a slightly lower performance, indicating that the additional features did not improve the model's predictive power.

Table S5

Confusion matrix of binary tree model with all features (cross-validation set)

		Predicted class	
	Model1 (N=377)	Unicorn	Non-Unicorn
True class	Unicorn	**28 (.88)**	4 (.39)
	Non-Unicorn	43 (.31)	**302 (.88)**
	Model2 (N=368)	Tortoise	Hare
True class	Tortoise	**109 (.62)**	68 (.38)
	Hare	64 (.38)	**104 (.62)**

Classification model selection

We evaluated multiple classification models during the model selection phase, utilizing both linear and non-linear algorithms. The performance of these models on the cross-validation set is presented in Table S6. For the first model, our objective was to detect unicorns, as they exhibited excellent learning performance. In this case, the repercussions of missing positive instances were considered to be more significant than the consequences of false positives. As a result, the selection of model 1 was based on both recall (to maximize the detection of unicorns) and accuracy. On the other hand, model 2 was developed to differentiate between hares and tortoises. In this scenario, our primary focus was on accuracy, aiming to correctly classify as many instances as possible. Therefore, the selection of model 2 was primarily based on accuracy. By considering the goals and characteristics of each model, we chose the RBF Support Vector Model for model 1 and the Random Forest model for model 2.

Table S6

Model Performance of Different Classification Algorithms (cross-validation set)

Model 1	Precision	Recall	F1 score	Mean Accuracy (SD)
Logistic Regression	.129	.562	.209	.681 (.050)
Decision Tree	.264	.594	.365	.869 (.044)
Random Forest	.568	.656	.609	.954 (.018)
Linear Support Vector Model	.129	.594	.212	.690 (.065)
RBF Support Vector Model	.311	.875	.459	.902 (.029)
Multiple Layer Perceptron	.286	.688	.404	.898 (.031)
Model 2				
Logistic Regression	.581	.525	.552	.556 (.041)
Decision Tree	.586	.576	.581	.580 (.066)
Random Forest	.648	.644	.646	.643 (.052)
Linear Support Vector Model	.604	.542	.571	.575 (.056)
RBF Support Vector Model	.594	.571	.582	.581 (.053)
Multiple Layer Perceptron	.588	.548	.567	.571 (.054)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1 score} = \frac{True\ Positive}{True\ Positive + \frac{1}{2}(False\ Positive + False\ Negative)}$$

Training conditions as predictors

Table S7

Model Performance After Adding Training Program Features (cross-validation set)

Model	Precision	Recall	F1 score	Mean Accuracy
Original model 1	.311	.875	.459	.902
model 1 + Training conditions	.358	.750	.485	.813
model 1 + Training conditions + Gamified/non-Gamified	.359	.719	.479	.800
Original model 2	.648	.644	.646	.643
model 2 + Training conditions	.664	.616	.639	.645
model 2 + Training conditions + Gamified/non-Gamified	.663	.644	.653	.649

Note. Training conditions represent 10 different algorithms.

Another approach to predict learning patterns

Even though our updating task that we used as a baseline measure was an untrained variant of the N-back task used during training, it might not capture the same baseline ability as the training task itself. Therefore, we considered using the initial perfor mance on the training task itself to predict learning patterns. Given that this was only one predictor, we applied a simple logistic regression model with exactly the same dataset. The confusion matrix is shown in Table S8. In the first model, the F1 score is .256, and in the second model, it is .565.

Table S8

Confusion matrix of the logistic model with one predictor (cross-validation set)

		Predicted class	
	Model1 (N=404)	Unicorn	Non-Unicorn
True class	Unicorn	**22 (.61)**	14 (.39)
	Non-Unicorn	114 (.31)	**254 (.69)**
	Model2 (N=368)	Tortoise	Hare
True class	Tortoise	**98 (.52)**	90 (.48)
	Hare	61 (.34)	**119 (.66)**

Cohort-based descriptions of learning pattern groups

Table S9 displays the frequencies of each learning pattern group by cohorts. The results indicate that there was no learning pattern specific to any particular cohort.

Table S9

Distribution of Learning Pattern Groups Across Cohorts

Cohort	Total N	Unicorn	Tortoise	Hare	Starfish	2-back algorithm	Outliers
1	266	25	103	114	1	11	12
2	138	7	57	43	27	/	4
3	164	13	76	68	3	/	4

Note. The "2-back algorithm" category represents participants from a specific algorithm that solely implemented the 2-back task. Therefore, this category was excluded from our analysis.

CHAPTER 2

Promoting Working Memory and Numeracy Skills in Kindergarteners: Optimal Intervention Designs for Children with Varying Levels of Math Skills

2.1 Introduction

Mathematics learning involves a range of domain-specific and domain-general skills. Even simple arithmetic problems require both knowledge about symbolic numbers and operations, as well as the cognitive resources to maintain information and manipulate problems. Specifically, in order to successfully solve a math problem, children have to rely on a variety of domain-specific skills, including non-symbolic approximate number skills (Fazio et al., 2014; Schneider et al., 2017) and symbolic number knowledge (Vanbinst & De Smedt, 2016), as well as domain-general cognitive executive function (EF) skills including working memory (WM; Bull & Lee, 2014, Geary & Hoard, 2005). Correlational evidence supports the reciprocal relationship between domain-specific and domain-general skills, and thus, to promote math performance, interventions have been targeting either domain-specific skills such as numerical knowledge, non-symbolic processing, and arithmetic fluency (DeWind & Brannon, 2012; Park et al., 2016; Sánchez-Pérez et al., 2018), or domain-general skills like WM (Bergman-Nutley & Klingberg, 2014; Söderqvist & Bergman, 2015; Ribner, 2020). A few studies have also combined these two approaches with the idea that this could maximize the training benefits (DePascale et al., 2024; Fuchs et al., 2020; Lee, 2023; Nemmi et al., 2016; Prager et al., 2023; Ribner and Libertus, 2021). However, no study to date has systematically compared the effectiveness of three cognitive intervention

types—domain-specific, domain-general, and combined training—in improving math and WM skills across varying baseline skills, which is the goal of the current paper. The exploration of combined training is critical, as it is expected to merge the benefits of both domain-specific and domain-general interventions yet remains untested.

In domain-specific training for young children, the focus typically lies on enhancing both non-symbolic and symbolic numerical skills, such as counting, number recognition, and symbolic comparison (e.g., Starr et al., 2013). These foundational skills are key indicators of later mathematical achievement, which is why those skills are targeted by interventions (Hawes et al., 2019; Lyons et al., 2014; Martin et al., 2014). Although these interventions are generally effective, their efficacy can vary with some children benefiting more than others (Reynvoet et al., 2021; Williams et al., 2022). Domain-specific training seems to be especially helpful for children who struggle with math, as it allows them to receive targeted instruction and practice in areas where they need the most support (Hassler et al., 2018). Contrary to this, Nelson and McMaster's (2019) meta-analysis study found that interventions were less effective for students at high risk of math difficulties and from low socio-economic backgrounds compared to those without such risks. Meanwhile, Chodura et al. (2015) found that the effectiveness of interventions did not vary in different levels of children's math learning difficulties. In addition to baseline performance in math, domain-general skills like EF can also affect intervention outcomes. For example, Dong et al. (2022) observed that EF skills are more positively associated with math skills in children with relatively lower mathematical skills. Furthermore, they found that mathematical interventions are particularly beneficial for children with deficits in both EF

and mathematical skills, highlighting the importance of considering both domain-specific and domain-general skills in mathematical interventions.

The effectiveness of domain-general training on math performance is still up for debate (Diamond & Ling, 2020; Kassai et al., 2019; Sasser et al., 2017), although some studies have shown promising effects (e.g., McClelland et al., 2019). For example, a study that targeted WM skills in kindergarteners found that the training enhanced participants' counting skills (Kroesbergen et al., 2014). Other studies have also shown significant improvements in numerical comparison skills and arithmetic fluency through WM training (Cowan & Powell, 2014). In Ramani et al.'s (2017) study, they implemented both domain-general and domain-specific training to improve children's mathematical skills. Their study found that both training approaches effectively increased number line estimation and numerical identification skills compared to a control group. However, in their later study, they observed limited benefits of domain-general training (Ramani et al., 2020). Interestingly, their most recent study showed significant math improvements after both WM training and combined WM+Math training conditions (DePascale et al., 2024). In contrast, other studies failed to observe improvements in math skills as a function of WM or EF training (Titz & Karbach, 2014, Wang et al., 2019), which is also reflected in a meta-analysis that reported small effect sizes on mathematical performance as a result of WM training (Schwaighofer et al., 2015). Given these mixed results, it is unclear whether domain-general training is a suitable approach to promote mathematical performance (Ribner & Libertus, 2021), and if so, for whom. The inconsistency in results may stem from ignoring individual differences, such as SES (socioeconomic status; e.g., Blair & Raver, 2014) and/or initial performance (McClelland et al., 2019). For example, Prager et al.

(2023) found that WM training specifically benefited individuals from lower SES backgrounds. This indicates the necessity of exploring intervention effects in larger samples, while carefully accounting for individual variances such as SES and initial skill levels.

Furthermore, while direct domain-specific training seems to reliably enhance math skills, indirect effects of domain-general training on math performance should also be considered. Specifically, a combination of domain-general and domain-specific approaches may maximize training outcomes (Nemmi et al., 2016; Munez et al., 2022; DePascale et al., 2024). Given that domain-general skills facilitate the acquisition of domain-specific skills (Geary, 2013), children might benefit most if they were to develop domain-general skills first, allowing them to acquire the necessary foundation upon which they can then build domain-specific skills (Clements et al., 2016; Fuhs et al., 2014; Schmitt et al., 2017). As such, an intervention that combines both domain-general and domain-specific approaches may provide a more comprehensive solution for individuals across a variety of skill levels.

Here, we conduct a secondary analysis of three intervention studies, which included domain-general training only conditions (WM training), domain-specific training only conditions (Numerical Knowledge (NK) training), and a combined training condition. By using the combined dataset from these three intervention studies, our analysis allows us to uncover the differential efficacy of domain-specific and/or domain-general interventions by relying on an expanded sample size that includes 469 kindergarten students. Specifically, we address the following research questions:

1. Which intervention training – WM, NK, or combined WM+NK – is the most effective in improving math and WM skills in kindergarteners, both in the short-term and long-term?
2. How do individual differences in baseline WM and math skills influence the effectiveness of these interventions, both in the short-term and long-term?

2.2 Methods

Participants

The current secondary analysis integrated data drawn from three studies (DePascale et al., 2024; Ramani et al., 2017, 2020) that included a total of 469 kindergarten children (aged between 4 to 10, *Mage* = 5.5, 49% female) who were distributed across race, ethnicity, as well as SES (see Table 2.1).

Table 2.1

Descriptive statistics of demographic information across the three studies

	Study 1			Study 2			Study 3			
Training Condition	WM	NK	Control	WM	NK	Control	WM	NK	WM+NK	Control
Sample size (N)	27	27	27	51	47	55	58	58	61	58
Female (%)	50	42	44	60	47	56	48	47	51	53
Mean age in months (SD)	72.1 (4.3)			72.6 (6.8)			70.1 (4.6)			
Mean SES (SD)*	0.0 (0.8)			0.3 (1.1)			-0.2 (0.7)			
Hispanic (%)	46.9			41.2			45.1			
African American or Black (%)	11.1			22.9			35.3			
Caucasian or White (%)	17.3			37.3			21.3			
Asian or Pacific Islander (%)	9.9			2.6			3.8			
American Indian or Alaska native (%)	0.0			1.3			0.4			
Biracial/Mixed race (%)	12.3			9.8			6.4			
Others (%)	0.0			2.0			1.7			
Unreported (%)	2.5			24.2			31.1			

Note. *SES (socioeconomic status) is a composite z-score of mother's education level (1 - Some High school coursework to 6 - Post graduate or Professional degree) and family annual income level (1 - Less than $15,000 to 7 - $151,000 or more).

In each of the three studies, participants were randomly assigned to a game-based training group or a control group. In a training group, participants received training in either NK or WM, or a combination of both (WM+NK). The training programs in the three

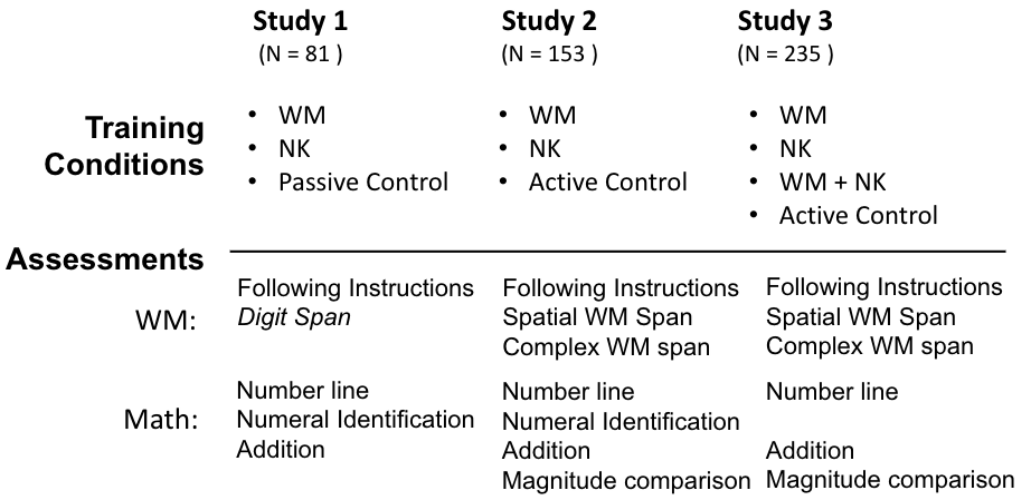
studies differed slightly in terms of visual themes used in the training games, but the training tasks themselves maintained a consistent structure. The primary results have been published in Ramani et al. (2017), Ramani et al. (2020), and DePascale et al. (2024), respectively.

Training Programs

The training procedures are described in previous works (Ramani et al., 2017; Ramani et al., 2020; DePascale et al., 2024). In short, all three studies consisted of 10 training sessions over the course of two weeks and consisted of two components: NK and WM training. Importantly, the third study also included a combined training condition, where participants first completed 5 sessions of WM training followed by 5 sessions of NK training. Before and after the training, participants' WM and math performance was assessed. Studies 2 and 3 incorporated a three-month post-intervention follow-up to evaluate the long-term effect of the training. Details are provided in the Figure 2.1.

Figure 2.1

Experimental design across the three studies



Note. WM: Working memory; NK: Numerical knowledge. The assessments were the same before and after training. Studies 2 and 3 also included long-term follow-up assessments. Digit Span (highlighted in italics) was the only task measured in Study 1.

Numerical Knowledge (NK) training: The same math training game was used in all three studies. Participants were presented with a 10 by 10 matrix of numbers from 1 to 100, with values increasing from left to right. The participants and the computer program controlled two characters, respectively. These two characters took turns to move along the matrix. During the turns controlled by the participants, they needed to press a spinner containing numbers from 1 to 6 and move the corresponding number of steps by tapping each space while counting along with the numbers on the matrix. During the computer-controlled turns, the character also moved a certain number of spaces from 1 to 6. In Studies 2 and 3, participants also answered a numerical magnitude question after every three rounds (e.g., which character is leading?).

Working Memory (WM) training: In this game, participants were shown a series of characters that were either right-side-up or upside-down. First, participants were asked to identify the orientation of each character. After all the characters were presented, participants had to recall the order by selecting the correct characters from a multiple-choice list.

Control groups: Study 1 included a passive control group where participants only received assessments within the same timeframe as the experimental group. In Study 2, a similar game as the NK training was used except that the numbered spaces were replaced with a repeating pattern of colors. In Study 3, participants were presented with two images side by side and asked to find differences between them by tapping on the screen to indicate where the differences in the pictures were located.

Assessments

Details for all assessments are provided in the original studies, however, a short description is provided in the following: Before and after the intervention, as well as several weeks after the posttest in Studies 2 and 3, participants completed several math and WM assessments. Different tasks were used as indicators of math and WM skills across the three studies, although some of the measures remained consistent. For the purposes of the current paper, we ensured that assessments were comparable by maintaining overlapping measures and excluding those that were specific to any single study. A composite z score was made of all tasks for each study that served as a dependent variable representing overall math or WM performance.

Math assessments: *Number line*: participants were presented with a number line from 0 (left) to 100 (right) and a target number above the number line (Ramani et al., 2017;

2020, DePascale et al., 2024). Participants were asked to mark the position of that target number on the line. The dependent variable was the accuracy of correct estimation calculated by the percentage of absolute error (PAE; Siegler & Booth, 2004). *Numeral identification:* Participants were asked to name Arabic numerals selected from 1 to 100. The dependent variable was the percentage of correct identification. *Addition:* Participants were asked to solve simple addition problems (presented problems were detailed in prior publications). The dependent variable was the total number of correctly solved problems. *Magnitude comparison:* Participants were shown pairs of numbers and asked to identify the larger of the two. The dependent variable was the percentage of trials with correct responses.

WM assessments: *Following instructions:* Participants were asked to follow a series of actions, like "put the green car into the blue box." The dependent variable was the total number of correctly performed actions (Gathercole et al., 2008; Ramani et al., 2017; 2020, DePascale et al., 2024). *Digit span*: Participants were asked to repeat a series of numbers (Wechsler, 1991). The dependent variables were the total number of trials recalled correctly. *Spatial WM Span*: Participants were asked to reproduce a series of spatial locations (Corsi, 1972). The dependent variable used was the total number of trials recalled correctly. *Complex WM Span*: Participants were asked to count a series of targets while ignoring distractors. At the end of each series, they were asked to recall their counts in the correct order. The dependent variable was the percentage of correctly recalled counts (Case & Goldberg, 1982; Ramani et al., 2020; DePascale et al., 2024).

2.3 Results

Preliminary Analyses

Descriptive statistics of demographic information across all studies are shown in Table 2.2. There were some differences in the population composition among the three studies. Study 2 and Study 3 had larger sample sizes in each training condition, which suggests they were better powered to detect differences between conditions. Participants in Study 3 were younger than those in Study 1 (*Cohen's d* = 0.45) and Study 2 (*Cohen's d* = 0.45), and their SES was lower compared to participants from Study 1 (*Cohen's d* = 0.34) and Study 2 (*Cohen's d* = 0.63). Studies 2 and 3 had a more diverse participant population compared to Study 1. Note though that Study 1 combined race and ethnicity into one question, making direct comparisons challenging. Study 3, compared to Study 2, had greater racial diversity, specifically, a higher proportion of African American or Black participants and a lower proportion of Caucasian/White participants (*Chi-Square test p* < .01). There were no differences observed in the gender distribution o f participants across the three studies (*Chi-Square test p* = .45). Despite these overall demographic variations, there were no differences within each study in terms of age, gender, SES, or race/ethnicity with respect to training conditions (all *ps* > .27). Additionally, completion rates were high across all three studies, with an average of 95% of the training sessions completed and over 90% of participants completing all posttests. This high level of completion suggests that the training and assessments were generally well-received.

To investigate the association between demographic factors and baseline math and WM abilities, we employed linear regression analyses using age in months, gender, SES, and race/ethnicity to account for variations in baseline performance. To make race and

ethnicity comparable across the three studies, we created two categories: Non-marginalized groups consisting of Caucasian/White, Asian, Biracial/Mixed race, or others (40.3%), and traditionally marginalized groups (53.5%), consisting of participants identifying as Hispanic, African American or Black, or American Indian or Alaska Native. The results revealed that age in months, gender, SES, and whether participants were from traditionally marginalized groups were all significant predictors of baseline math abilities: older children scored higher in math skills ($p < .001$); boys scored higher than girls ($p = .001$); children from higher SES-backgrounds demonstrated higher math skills ($p < .001$); and children from traditionally marginalized groups scored lower ($p = .014$), all after controlling for other demographic factors. Additionally, SES and whether participants were from traditionally marginalized groups were significant predictors of baseline WM abilities, with higher SES levels correlating with higher baseline WM performance ($p = .005$) and traditionally marginalized groups showing lower baseline WM performance ($p = .005$). Details are shown in Table 2.2.

Table 2.2

Regression results for baseline performance

Outcome Measure	Math baseline	WM baseline
Predictors		
Age in months	.03 (.01)**	.01 (.01)
Female	-.31 (.09)**	-.01 (.09)
SES	.22 (.05)**	.16 (.06)*
Marginalized status	-.24 (.10)*	-.28 (.10)*
Adjusted R^2	.14	.06
Observations	429	429

Note. The reported values are standardized coefficients with standard errors in parentheses. The non-marginalized group served as the reference group in the analysis. *p < .05; **p <=.001

To evaluate the consistency of our results across the three studies, we applied an identical regression model to each study separately. For baseline math skills, SES emerged as a significant predictor in Studies 2 and 3. Being in a traditionally marginalized group, as opposed to a non-marginalized group, only showed significance in Study 2. Age, measured in months, was only significant in Study 3, as was the case for gender. For baseline WM skills, the influence of SES approached significance in Study 1 and reached significance in

Study 3. Traditionally marginalized groups' status was driven by Study 3. Collectively, these findings suggest that SES was a consistent factor in influencing baseline math and WM skills. The other factors tend to show group differences in baseline performance, but critically, they only minimally contribute to the variance. Still, their potential impact on training efficiency cannot be completely disregarded. Following evidence from previous research that indicates SES moderating effect (e.g., Prager et al., 2023) and gender differences in getting benefit from mathematical training (Lee et al., 2021), SES and gender included as covariates in subsequent analyses. For completeness, the effects of the other factors are explored and presented in the Supplementary Materials.

Differential Outcomes as a Function of Training Condition

To better quantify and compare the effectiveness of different training conditions and understand the general training effects regardless of demographics, we calculated the Net effect size (d) for each treatment group within each study. Net effect size here refers to the difference in performance between the experimental group and the control group from pretest to posttest, or from pretest to follow-up test, respectively. This can be expressed as:

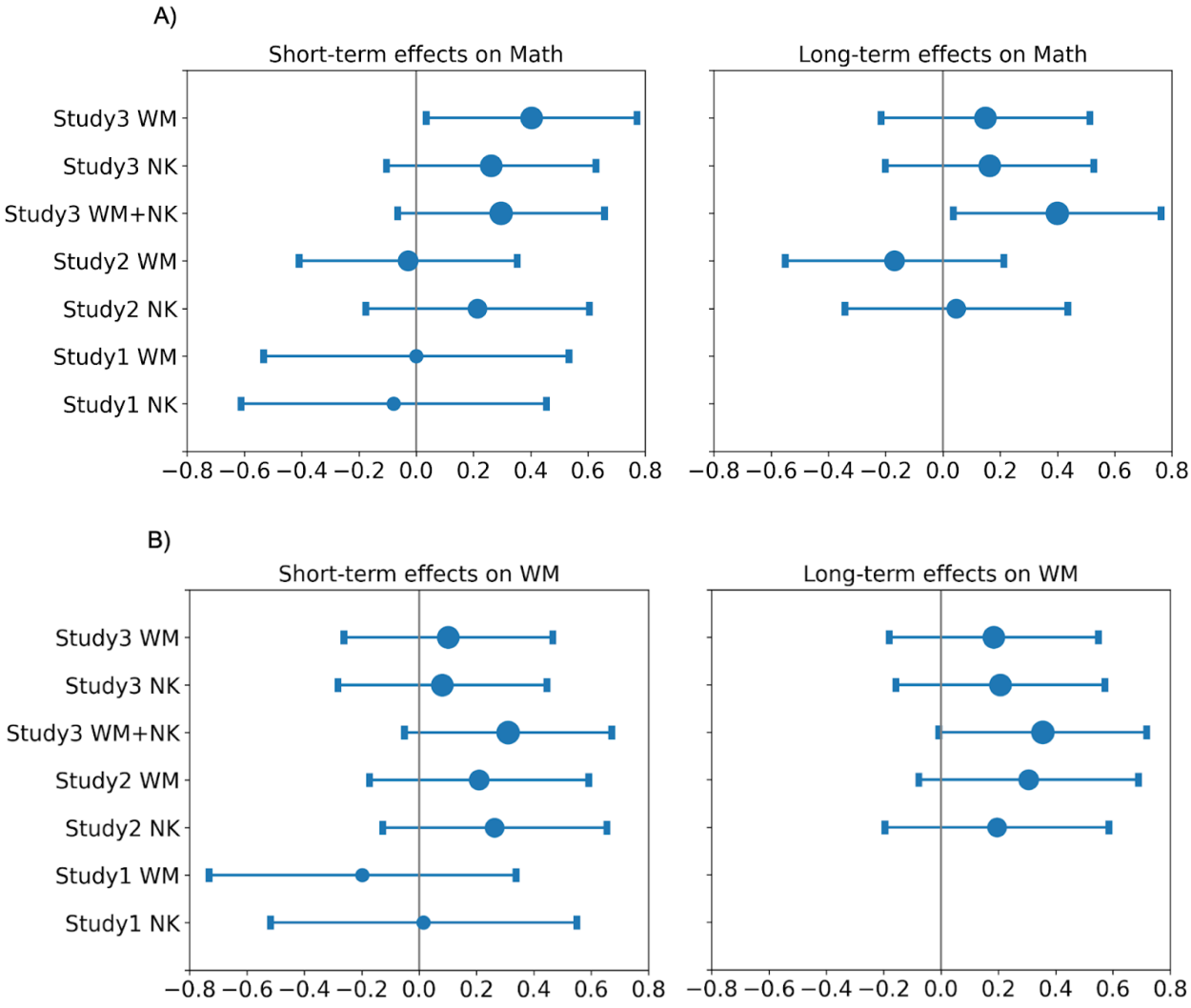
Net effect size = ($Posttest - Pretest$ of the $training\ group$) − ($Posttest - Pretest$ of the $control$ group)$Pooled$ Standard Deviation $of\ the\ difference$

As such, we considered both short-term and long-term training-related outcomes on math and WM. As shown in Figure 2.2, the WM-only intervention in Study 3 yielded the most pronounced short-term gains on math, with a medium effect size (*Net effect size* = .40, *95% CI* = .37). However, the combined WM+NK intervention demonstrated the greatest Net effect size (d =.40, *95% CI* = .36) in long-term math gains compared to the other

interventions. Notably, the combined intervention also showed an increase in math outcomes at the follow-up (*Net effect size* increased from .30 to .40), in contrast to the other groups who showed a decline at follow-up. Although no significant effects were observed in WM outcomes, the combined intervention group showed the largest effect sizes compared with the other groups (short-term *Net effect size* = .31, *95% CI* = .36; long-term *Net effect size* = .35, *95% CI* = .36).

Figure 2.2

Net effect sizes of training groups in each study



Note. Net effect sizes of training groups on A) Math outcomes; B) WM outcomes. X-axes are Net effect sizes; Y-axes are the training groups under each study. Dot sizes indicate the sample size of each training group, and error bars represent 95% CIs.

Differential Math Outcomes as a Function of Baseline Performance

To investigate whether the effects of the interventions varied by baseline skills, we conducted interaction analyses to evaluate the intervention effects as a function of pretest scores (Intervention Type by Baseline Score). We conducted four regression models to assess both short-term and long-term effects—these encompassed measures of posttest performance, short-term gains (pre vs posttest), follow-up test performance, and long-term gains (pre vs. follow-up). The models incorporated baseline WM and math skills, as well as intervention type as predictors, along with interaction terms. To mitigate inter-study variability, we introduced 'Studies' as a random effect in the mixed linear regression model. As described in preliminary analysis, SES and gender were included as covariates in the mixed model.

As reported in Table 2.3, there was a positive correlation between baseline math performance and math performance in both the posttest and follow-up tests. However, baseline math performance was negatively correlated with both short-term and long-term math gains. This suggests that participants with initially lower math skills benefited more from the interventions, yet they still underperformed compared to their peers with higher initial skills. Additionally, baseline WM performance tended to correlate with follow-up math performance and long-term improvements, suggesting a potential role for WM in ongoing math learning (both *p-values* = .067). Notably, the analysis revealed a near-significant interaction between baseline math performance and the intervention type—specifically, NK-only versus WM-only interventions—influencing long-term math performance and training gains (p = .089). As well as a significant interaction between baseline math performance and NK-only versus combined WM+NK intervention (p = .004).

These results indicate that participants with relatively lower baseline math skills tended to benefit more from NK-only training in the long run. In contrast, those with relatively higher baseline math skills seemed to benefit more from the combined WM+NK training. The interaction effects between Intervention Type and Baseline Score persisted even after accounting for both, SES and gender differences.

Table 2.3

Regression results for mathematics outcomes

Outcome Measure	Posttest	Short-term gains	Follow-up test	Long-term gains
Predictors				
Baseline WM	.06 (.05)	.13 (.10)	.09 (.05)†	.18 (.10)†
Baseline Math	.86 (.05)***	-.27 (.10) **	.80 (.05)***	-.32 (.10) **
NK training	.04 (.06)	.07 (.12)	.08 (.06)	.16 (.12)
Combined training	-.01(.07)	-.02 (.15)	.12 (.07)	.23 (.14)
Interaction terms				
Baseline WM×NK	.01 (.06)	.02 (.14)	-.02 (.07)	-.05 (.14)
Baseline Math×NK	-.08 (.07)	-.19 (.14)	-.12 (.07)†	-.24 (.14)†
Baseline WM × Combined	-.02 (.09)	-.04 (.19)	-.06 (.09)	-.13 (.18)
Baseline Math × Combined	.02 (.08)	.05 (.19)	.12 (.08)	.24 (.16)
Covariants				
SES	.10 (.03)*	.21 (.06)*	.04 (.03)	.09 (.06)
Female	-.03 (.05)	-.07 (.11)	-.10 (.06)†	-.21 (.11)†
Random effects				
Studies	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)
Observations	305	305	254	254

Note. WM training served as the reference group. Standardized beta coefficients. Standard errors in parentheses. † p<.09, * p<0.05, ** p<0.01, *** p<0.001

Differential WM Outcomes as a Function of Baseline Performance

Similar findings emerged when examining baseline WM skill's main effects on WM outcomes (see Table 2.4), with those showing lower initial WM performance experiencing greater benefits from the intervention. Conversely, participants with higher baseline math skills demonstrated more pronounced WM improvements, both in short-term and long-term outcomes. Again, the interaction between baseline math performance and the intervention type - combined intervention versus WM-only intervention - did only approach significance, but it was consistent for both follow-up tests and long-term gains (both *p-values* = .062). Specifically, participants with high baseline math skills tended to show the most substantial long-term improvements in WM skills when undergoing combined training, outperforming those who received WM-only training. While participants with lower baseline math skills tended to improve more in WM-only intervention condition. This suggests that the combined training approach might be particularly beneficial for individuals with higher initial math skills in enhancing long-term WM skills.

Table 2.4

Regression results for WM outcomes

Outcome Measure	Posttest	Short-term gains	Follow-up test	Long-term gains
Predictors				
Baseline WM	.31 (.08)***	-.73 (.08)***	.28 (.09)**	-.73 (.09)***
Baseline Math	.29 (.08)***	.32 (.09)***	.29 (.09)**	.30 (.09)**
NK training	.08 (.10)	.09 (.11)	-.08 (.11)	-.08 (.11)
Combined training	.18 (.13)	.19 (.13)	.08 (.13)	.09 (.13)
Interaction terms				
Baseline WM×NK	.03 (.11)	.04 (.12)	.02 (.12)	.02 (.13)
Baseline Math×NK	.03 (.11)	.03 (.12)	.09 (.12)	.10 (.13)
Baseline WM × Combined	-.04 (.15)	-.04 (.17)	-.03 (.16)	-.04 (.16)
Baseline Math × Combined	.24 (.14)	.25 (.15)	.26 (.14)†	.27 (.15)†
Covariants				
SES	.11 (.05)*	.12 (.06)*	.05 (.06)	.05 (.06)
Female	.16 (.09)†	.17 (.10)†	.10 (.10)	.10 (.10)
Random effects				
Studies	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)
Observations	306	306	254	254

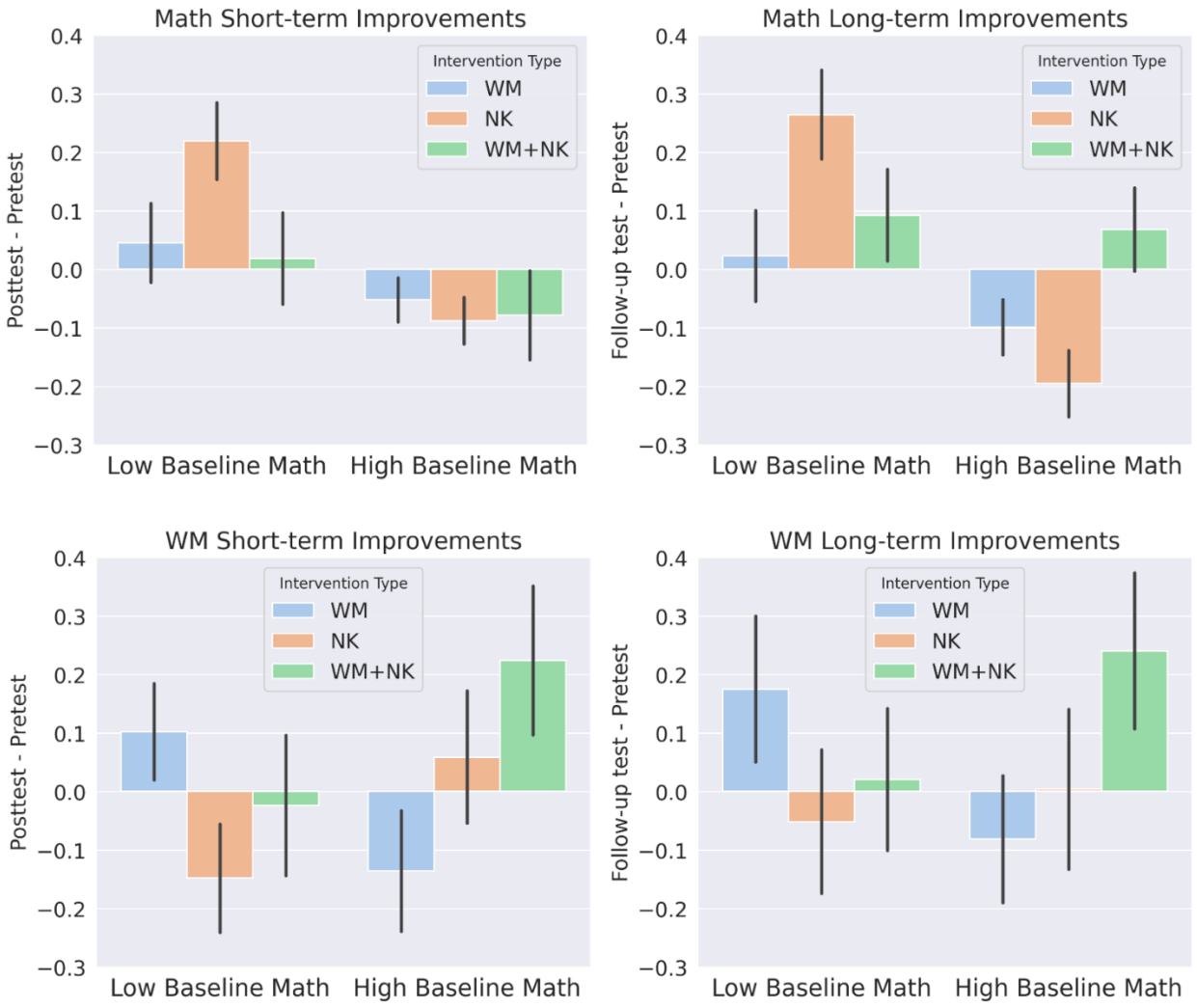
Note. WM training served as the reference group. Standardized beta coefficients. Standard errors in parentheses. † p<.09, * p<0.05, ** p<0.01, *** p<0.001

Training Improvements by Initial Skill Groupings

To further explore how baseline skills might influence intervention outcomes, participants across all interventions were divided into two groups based on their baseline math skills: the low group, comprising those participants who scored in the lowest 33%, and the high group, consisting of those participants who scored in the highest 33%. These findings of our analyses are detailed in Figure 2.3: Participants with low baseline math performance showed more pronounced improvements in math skills following NK-only training compared to combined WM+NK training (short-term d = 0.50, long-term d = 0.41) and WM-only training (short-term d = 0.41, long-term d = 0.54). However, for participants with high baseline math performance, the combined training was more effective for long-term improvement (d = 0.75 compared to NK-only training, and d = 0.54 compared to WM-only training). Regarding WM outcomes, WM-only training was most effective for the low baseline math performance group (d = 0.46 compared to NK-only training, and d = 0.21 compared to combined training). In contrast, for the high baseline math performance group, the combined training approach was more beneficial (d = 0.30 compared to NK-only training, and d = 0.48 compared to WM-only training). In contrast, grouping participants based on baseline WM skills yielded no significant differences in WM and math outcomes as a function of intervention types.

Figure 2.3

WM and math improvements as a function of baseline math skills and training types



Note. Sample sizes for each group were shown in Supplemental Materials (Table S5). Error bars represent standard errors of the mean.

2.4 Discussion

The present study aimed to uncover the potential differential effects of domain-specific and domain-general training games on math and WM learning outcomes by relying

on combined data from three separate studies. Critically, this larger dataset allowed us to explore the potential impact of individual differences in math and WM skills at baseline, and the extent to which those baseline skills differentially impact the intervention benefits. Our findings suggest that an intervention incorporating both WM and NK training generally yielded the most substantial improvements in math learning. The effectiveness of the combined training is consistent with previous work that emphasizes the importance of promoting both, domain-specific and domain-general skills (Clements et al., 2016, Fuhs et al., 2010, 2014). WM training generally targets overarching domain-general cognitive functions like attention, adaptability, and processing speed. Numerical training tends to focus on domain-specific mathematical knowledge and techniques, which also require domain-general skills to complete various math tasks. By first building WM capacity and resources, subsequent numerical training can be facilitated, as more cognitive resources will be available for learning math once the scaffold of WM is established. Additionally, combined training likely offers a broader array of cognitive strategies for solving mathematical problems. This holistic approach may provide a more comprehensive skill set for problem-solving and adapting to new mathematical challenges (DePascale et al., 2024). Given that combined training provides the opportunity to integrate both domain-general and domain-specific skills, we hypothesized that WM capacity would be reinforced and improved through its application to math tasks as well. However, our study did not yield significant effects on WM outcomes across all intervention types. This lack of impact may be attributable to the relatively short duration of the intervention, comprising merely 10 sessions. This finding contrasts with the results reported in Zhang et al. (2019), where using 20 sessions consistently produced notable effects.

It is also noteworthy that for both math and WM, individuals with lower initial math skills benefited more from single-type (domain-specific) training, whereas those with higher initial skills benefited from combined training. This pattern aligns with earlier research suggesting that domain-specific training particularly aids those starting from a lower skill level (Hassler et al., 2018). This finding implies that while domain-general skills (e.g. EF) are recognized for its role in the early stages of math learning, its effectiveness may be limited without a foundational numerical knowledge. However, once a basic numerical understanding is established, domain-general training becomes more valuable. We found that combined training is most effective for individuals with higher baseline performance in math. This might be because their pre-existing strong skills make single-type interventions less relevant or engaging for students, as evidenced by their performance in posttests or follow-up tests, which can be lower than in pretests with a singular intervention (Figure 2.3; high baseline math groups showing negative values on performance changes). Yet, a prior study by Dong et al. (2022) did not observe variations in the effectiveness of combined training based on participants' baseline EF and math skills. It's important to note, though, that their study's approach to combined training involved integrating EF exercises within the existing curriculum, rather than the addition of direct domain-general training. This distinction underscores the need for further exploration into the heterogeneous effects of combined training.

Although our study demonstrated the direction of the interaction between treatment type and baseline performance, our findings should be considered preliminary, given that our effect sizes were relatively low. This could be due to our sample size not being sufficiently powered to explore interaction effects robustly. Additionally, merging

three studies may have introduced more noise into the data, making it more challenging to detect significant effects.

Nonetheless, these preliminary insights hold substantial value for educational researchers and practitioners focusing on mathematical interventions and support strategies for students struggling with math. Overall, our data suggest that a multi-faceted approach to cognitive training, encompassing various skill areas, is more advantageous than programs targeting a singular cognitive capacity. However, it is important for the design of educational interventions to take into account students' individual math skills to tailor the intervention effectively. For students with lower baseline skills, interventions focusing on a specific area might yield the most benefits. Conversely, students with higher initial math skills might experience greater gains from interventions that combine both domain-general and domain-specific skills.

Given the potential benefit of combined domain-general and domain-specific training, future programs could adopt a more synergistic approach. For example, naturally integrating WM into math tasks so that participants need to use both their WM and numerical skills simultaneously to complete a task. Besides the training content, future empirical work should concentrate on establishing the optimal parameters for such training programs, including their duration and intensity, and should be adjusted according to the participant's demographics and proficiency in math skills. Additionally, it is crucial to identify which math tasks most accurately reflect the efficacy of the interventions. Further studies are also warranted to examine the sustainability of the gains achieved through these training efforts and their impact on other academic subjects (e.g., science learning). Such research is essential to understand the broader applicability of cognitive training in

educational settings and its potential to foster academic excellence across various disciplines.

Supplementary Materials of Chapter 2

Table S1

Regression results for Math outcomes with all covariates included

Outcome Measure	Posttest	Short-term gains	Follow-up test	Long-term gains
Predictors				
Baseline WM	.06 (.05)	.12 (.10)	.09 (.05)†	.18 (.10)†
Baseline Math	.85 (.05)***	-.30 (.10)**	.80 (.05)***	-.32 (.10)**
NK training	.03 (.06)	.07 (.12)	.08 (.06)	.16 (.12)
Combined training	.00 (.07)	.01 (.15)	.11 (.07)	.22 (.15)
Interaction terms				
Baseline WM×NK	.01 (.06)	.03 (.14)	-.03 (.07)	-.05 (.14)
Baseline Math×NK	-.09 (.07)	-.20 (.14)	-.11 (.07)	-.23 (.14)
Baseline WM × Combined	-.03 (.09)	-.06 (.19)	-.06 (.09)	-.12 (.18)
Baseline Math × Combined	.01 (.08)	.03 (.18)	.12 (.08)	.25 (.16)
Covariates				
SES	.09 (.03)**	.20 (.07)**	.04 (.03)	.09 (.07)
Female	-.03 (.06)	-.06 (.11)	-.10 (.06)†	-.20 (.11)†
Marginalized status	-.04 (.06)	-.07 (.12)	-.01 (.06)	-.01 (.12)
Age in Month	.01 (.01)	.02 (.01)	-.00 (.01)	-.01 (.01)
Random effects				
Studies	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)
Observations	302	302	251	251

Note. WM training served as the reference group. Standardized beta coefficients. Standard errors in parentheses. † p<.09, * p<0.05, ** p<0.01, *** p<0.001

Table S2
Regression results for WM outcomes

Outcome Measure	Posttest	Short-term gains	Follow-up test	Long-term gains
Predictors				
Baseline WM	.30 (.08)***	-.74 (.08)***	.28 (.09)**	-.73 (.09)***
Baseline Math	.28 (.08)**	.30 (.09)**	.29 (.09)**	.30 (.09)**
NK training	.08 (.10)	.08 (.11)	-.07 (.11)	-.07 (.11)
Combined training	.19 (.13)	.20 (.14)	.09 (.13)	.09 (.13)
Interaction terms				
Baseline WM×NK	.04 (.11)	.05 (.12)	.02 (.12)	.02 (.13)
Baseline Math×NK	.02 (.11)	.03 (.12)	.10 (.12)	.10 (.13)
Baseline WM × Combined	-.04 (.15)	-.05 (.17)	-.04 (.16)	-.04 (.16)
Baseline Math × Combined	.23 (.14)	.25 (.16)	.27 (.15)†	.28 (.15)†
Covariates				
SES	.08 (.06)	.09 (.06)	.02 (.06)	.02 (.06)
Female	.17 (.09)†	.18 (.10)†	.10 (.10)	.11 (.10)
Marginalized status	-.13 (.10)	-.14 (.11)	-.16 (.11)	-.16 (.11)
Age in Month	.01 (.01)	.01 (.01)	-.01 (.01)	-.01 (.01)
Random effects				
Studies	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)	<.001 (<.001)
Observations	303	303	251	251

Note. WM training served as the reference group. Standardized beta coefficients. Standard errors in parentheses. † p<.09, * p<0.05, ** p<0.01, *** p<0.001

CHAPTER 3

Forecasting Cognitive Training Improvement from Early Learning Performance

Committee chair Susanne M. Jaeggi has an indirect financial conflict of interest with MIND research institute, whose interests are related to this work.

3.1 Introduction

Over the past decade, the effectiveness of cognitive training to improve cognitive skills has been a major focus of interest and debate (Au et al., 2016; Green & Newcombe, 2020; Karbach & Unger, 2014; Melby-Lervåg et al., 2016; Pahor et al., 2022; Redick, 2019; Simons et al., 2016). One of the key challenges in this field is the considerable variability in training gains among individuals, both during the intervention programs and in pre- and post-intervention assessment tasks (Flak et al., 2019; Holmes et al., 2009; Karbach et al., 2015; Katz et al., 2021; Meiran et al., 2019; Redick, 2019; Wiemers et al., 2019). These individual differences can be attributed to several factors (Feng et al., 2023; Jaeggi et al., 2014; Ørskov et al., 2021), including baseline cognitive abilities, personal characteristics, environmental factors (as discussed in Chapter 1), the design of optimal training programs (as discussed in Chapter 2), and the overall suitability of the training regimen for each participant (Harrell et al., 2021). Even if a training program has shown to be effective for a particular population, it may be more beneficial for some individuals while not providing sufficient benefit for others, especially if the training regimen is not well-suited to their needs. Therefore, it is essential to monitor each individual's progress to assess whether the current regimen is yielding beneficial results. In practical cognitive training programs, this

ongoing evaluation helps determine if adjustments are needed to maximize the effectiveness of the training.

The current study focuses on improving children's cognitive skills especially their math fluency as early cognitive and math skills could predict long-term learning outcomes (e.g., Alloway, & Alloway, 2010; Bull et al., 2008; Eason et al., 2022; Gathercole et al., 2003; Geary, 2011; Melby-Lervåg & Hulme, 2013; Passolunghi et al., 2008). In developing an optimal training program, the combination of domain-general and domain-specific approaches may maximize training outcomes (DePascale et al., 2024; Munez et al., 2022; Nemmi et al., 2016). Previous research found that domain-general skills facilitate the acquisition of domain-specific skills (Geary, 2013), children might benefit most if they were to develop domain-general skills first, allowing them to acquire the necessary foundation upon which they can then build domain-specific skills (Clements et al., 2016; Fuhs et al., 2014; Schmitt et al., 2017). As such, our recent study designed an intervention that first trained kindergarteners' domain-general skills and then domain-specific skills, showing significant improvements in math performance (Chapter 2). Building on these findings, we propose that integrating domain-general skills within domain-specific tasks could provide a more comprehensive training approach. This method would simultaneously train both skill sets, maximizing the effectiveness of the training time.

Thus, for the current study, we designed a novel online training program that synergistically combines executive function and math fluency to maximize training effects on primary students, building on previous findings (DePascale et al., 2024; Chapter 2). Besides examining whether the training program's efficacy aligns with our Chapter 2 findings, the primary goal is to take a further step by monitoring students' training gains

through predictive models. Specifically, the prediction model aims to forecast individuals' training gains as early as possible. Early prediction can facilitate timely interventions, allowing for personalized adjustments that optimize the learning experience for each participant. This approach is crucial for identifying individuals who may struggle and need additional support, ensuring they fully benefit from the training program. This study served as a test-bed to evaluate the feasibility of such an approach.

Participating in an online training program and adhering to a training regimen can be challenging (Tullo et al., 2024), especially for elementary school students (Forbes et al., 2018; Lachheb et al., 2021). Previous studies have often observed that some individuals fail to show improvement in their performance over several weeks of training (Flak et al., 2019; Holmes et al., 2009; Karbach et al., 2015). This lack of progress could be due to unsuitable training content or a loss of motivation during repetitive training, leading to disengagement (Lumsden et al., 2016; Vermeir et al., 2020). Typically, researchers detect disengaged individuals only after analyzing the data once all participants have completed the training. Therefore, it would be extremely valuable to identify those who may need additional support early on to enhance their engagement and ensure they fully benefit from the training program.

Studies exploring the early prediction of training gains before or during training are scarce. Previous research has predominantly focused on describing training trajectories using linear growth models or conducting correlational studies to link existing individual factors with changes in assessment performance between pretest and posttest (e.g., Bürki et al., 2014; Guye et al., 2017; Karbach et al., 2017). In contrast, our recent research utilized pre-training personal cognitive skills and characteristic variables as predictors to forecast

individual learning patterns during training (Feng et al., 2024; Chapter 1). This recent research provides insight into the feasibility of predicting training gains beforehand. Building on this approach, the current study plans to use individuals' early training performance as predictors to forecast training gains. This method aims to enhance previous prediction models by incorporating new predictive factors to better anticipate individual performance, as performance during training may better reflect their interaction with the training program than existing cognitive profiles.

To date, very few studies have used participants' past training performance data to predict their subsequent performance. For example, Sandeep et al. (2020) used Bayesian filtering approaches to model participants' training trajectories, aiming to adjust the difficulty of subsequent sessions based on past training performance. This type of prediction is widely used in industry, such as Duolingo's use of spaced learning algorithms (Settles & Meeder, 2016) to decide the next word for review. However, these predictive models generally focus on forecasting the next specific response rather than the overall learning progress of the participants. Predictive models that can provide an early indication of whether participants are still on the right track are needed, and ultimately, allowing for adjustments in the training plan, pausing the training, or introducing new motivational interventions. This early prediction of cognitive training gains is akin to recent diagnostic prediction research, where machine learning offers promising tools for more accurate and early predictions of cognitive impairment in the aging population to facilitate early intervention (Ansart et al., 2021; Aschwanden et al., 2020; Grueso & Viejo-Sobera, 2021). Similarly, our early prediction of cognitive training gains plans to utilize machine learning, aiming to provide insights into the optimal timing for assessing potential

training benefits, enabling us to offer additional assistance early in the intervention process.

In sum, this study focuses on three key research questions to address the research gap in the early detection of cognitive training gains:

1. Does integrated executive function and math training improve participants' performance on executive function and math tasks?
2. To what extent can early learning performance predict overall training improvements?
3. How does participants' performance vary during early learning sessions?

To investigate these questions, we conducted a three-week intervention with 4th and 5th graders, combining executive function and math fact fluency training. Using machine learning models, we predicted learning outcomes based on early session performance. Additionally, cluster analysis was applied to categorize participants' early learning performance into distinct groups to better understand the information provided by early learning performance.

3.2 Methods

Participants

Participants were recruited from 10 schools in the Attleboro School District in Massachusetts and the Merced, Lowell, Fountain Valley, Santa Ana, and Ocean View School Districts in southern California. Data collection comprised three studies. The first study was a pilot to test the administration of our training program. The second study added more executive function and math measures, and the third study implemented additional executive function assessment tasks. Despite these variations, all three studies used the

same training program, prompting us to merge the datasets for the current study. A total of 325 grade 4 and grade 5 students (17 classrooms) participated in the intervention. The average age of the students was 9.86 years (SD = 0.74). Of these students, 37% identified as girls, 40% as boys, and 23% preferred not to say or did not disclose their gender. The ethnic composition was 38% Hispanic, 28% White, 7% Black, and 11% Asian. Given that the purpose of this study was to predict performance during training, 232 additional students (11 classrooms) who served as a business-as-usual control group were not included in the prediction analysis but were used as a reference group to examine training efficiency. An in-depth comparison between the two groups will be discussed in another paper. Before data collection, families were provided with an opt-out consent form that only had to be returned if caregivers did not want their children to participate. The study procedures were approved by the Institutional Review Boards at the University of California, Irvine, and the University of Maryland at College Park.

Intervention Group

Students primarily used our novel training platform, ST Math Fluency+, which combined executive function training with math fact fluency. The platform included seven different games, but for the current training period, students engaged with three specific games. Students were instructed to train on at least four different days (sessions) per week using Fluency+ on their personal or their parents' phones, without any administrative supervision. During each session, students could choose between two presented games to start with and played for a total of 15 minutes. They cycled through the games to ensure equal playtime for each game. All games focused on math fact fluency and incorporated executive function training (see Figure 3.1).

Light Beam. A light beam was blocked by shapes that were then covered by locks. To unlock the shapes, participants needed to solve math fact problems and subsequently recognize the shapes. This game required both math skills and working memory.

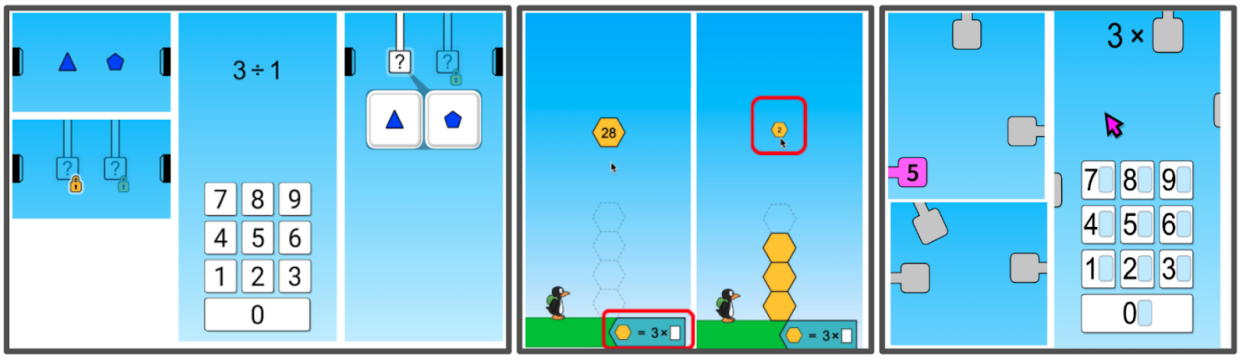
Hex Climber. This game targeted inhibitory control. Students decided if a falling hex-shaped number could be divided by a given number at the bottom right of the screen. For example, with the number 28 and a divisor of 3, students had to decide if 28 was divisible by 3. If it was, they let the hex fall; if not, they tapped it to keep it and build a hex bridge.

Paddle Shifter. In this game, students saw paddles, each representing a number. After viewing the numbers, the paddles were flipped over and moved around the screen. Students needed to recall a hidden number from one of the paddles and use it in a math fact calculation. This game targeted multitasking and math skills.

Each game adjusted its difficulty based on student performance, becoming more challenging with success and easier with struggle. The math facts that were harder for students to solve appeared more frequently. Thus, the difficulty level achieved in a session represents a student's performance, serving as the dependent variable. To better represent the overall training performance from the three Fluency+ games (Light Beam, Paddle Shifter, and Hex Climber), we computed z-scores of the difficulty level. For example, the z-score of participant A's performance in session 1 was calculated as follows: Participant A's raw score of session 1 minus the mean score across all sessions in all participants, divided by the SD across all sessions in all participants. We then averaged the z-scores across the three games for each session. Figure 3.2 illustrated the average learning curve based on aggregated data of all participants.

Figure 3.1

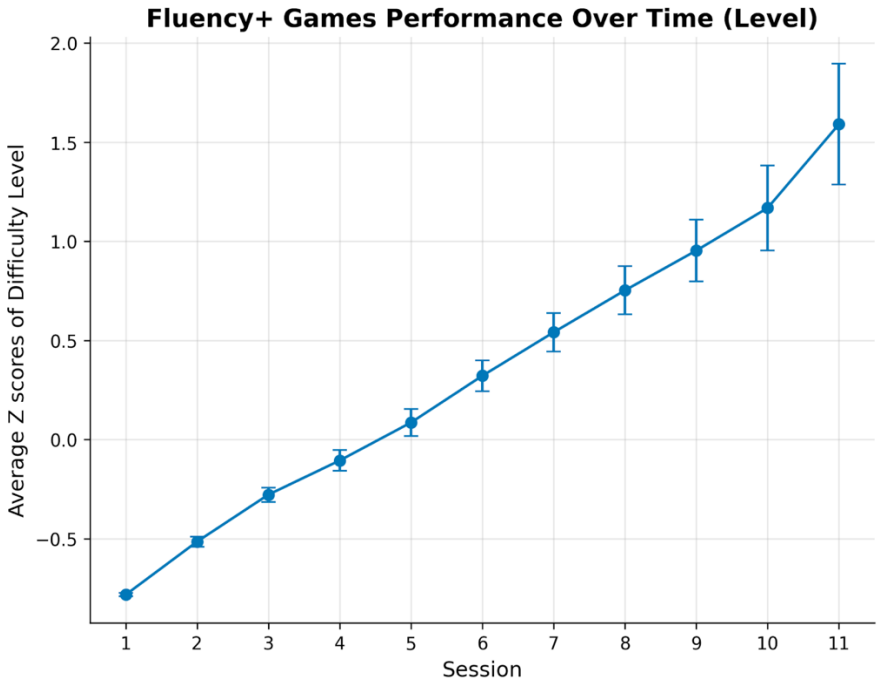
Three Executive Function + Math Games used in Training Program



Light Beam Hex Climber Paddle Shifter

Figure 3.2

Aggregated Learning Trajectory



Note. Participants' average Z scores of levels across training sessions. Error bars represent standard error of the mean.

Control Group

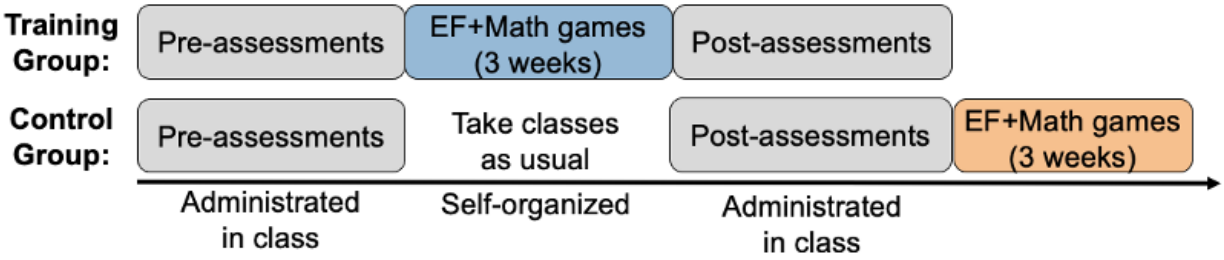
A business-as-usual control group served as a comparison for the intervention group. To minimize disruption, we did not ask the control group's teachers to focus on any specific content. Instead, we requested that they continue with their originally planned curriculum.

Procedure

Students in the intervention group were tested on various outcome measures before and after the intervention, while the business-as-usual control group was tested at the same intervals (see Figure 3.3). To ensure the training games remained potentially beneficial to the control group, they were introduced to the game after the post-test. The test administration was identical at both pre-test and post-test. Typically, students were tested over two different days (two days for pre-test and two days for post-test), with each test day taking approximately 60 minutes to complete. On one test day, students used school-provided Chromebooks to complete surveys, including a mindset survey (Petscher et al., 2021), a newly developed math-related executive function survey (Feng et al., 2022), and a fraction test. At post-test only, a demographics survey was also administered. On the other test day, students used Amazon Fire tablets provided by the research team to complete app-based tests (see below for outcome measures).

Figure 3.3

Training Procedure



Outcome Measures

As mentioned above, data collection comprised three studies that shared the same training program but had different outcome measures. To combine the three datasets, we included only the overlapping outcome measures: the fraction test and math fluency test for math skills, and the working memory span test for working memory skills. Since all participants took the same assessments before and after the intervention, we calculated the z-scores of gain scores to represent their assessment gains for both math and working memory measures. Specifically, within each dataset, we calculated the z-score of the posttest score minus the pretest score for participants (including both control and intervention groups). To better represent participants' overall math skills, we calculated the average z-score from the two math measures. The working memory span performance served as the measure of working memory skills.

Math Measures

Fraction Test. This self-developed measure consisted of 16 items with various math problems involving fractions, administered on a computer using Qualtrics. Test construction was guided by the Common Core fraction standards for fourth grade. The problems came in a variety of forms, such as simple fraction problems or text problems

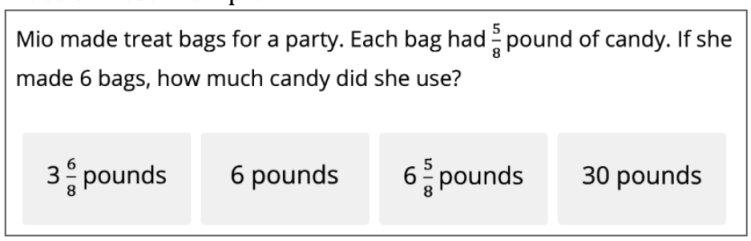
(see Figure 3.4). Half of the items were procedural in nature, and the other half were conceptual. Each item included either three or four answer alternatives, one of which was correct. The same test was administered at both pre- and post-assessment; however, the order of the problems and the answer alternatives were randomized. Students were given a maximum of 30 minutes to complete the test. The dependent variable was the number of correctly answered problems.

Math Fact Fluency Test. Based on the work by Gliksman et al. (2022), students were presented with a series of math facts that they were instructed to solve as quickly as possible without making an error. The test was administered on a tablet, where the math facts were shown on one side and a number pad on the other side was used to submit the answer. The test was divided into two parts, each lasting three minutes. The problems included single and double-digit addition, subtraction, and multiplication (no division). The dependent variable was the number of correctly answered problems.

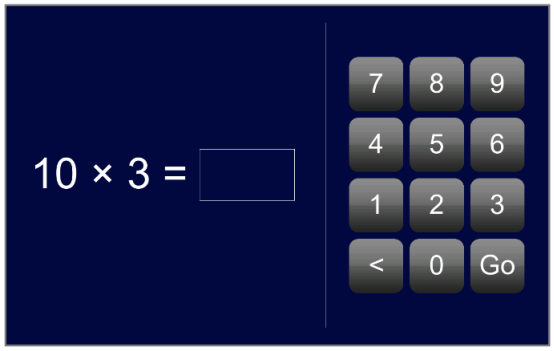
Figure 3.4

Math Measures

Fraction Test Example



Math Fact Fluency Test Example



Working Memory Measure

Simple Span. In this task, inspired by the Corsi-Block Tapping task (e.g., Kessels et al., 2000), students were presented with sequences of creatures that appeared in different locations on the screen (Ramani et al., 2017). Students were asked to repeat the presented sequence which increased in length after successful repetition (see Figure 3.5).

Figure 3.5

Working Memory Measure

Working Memory Span Test



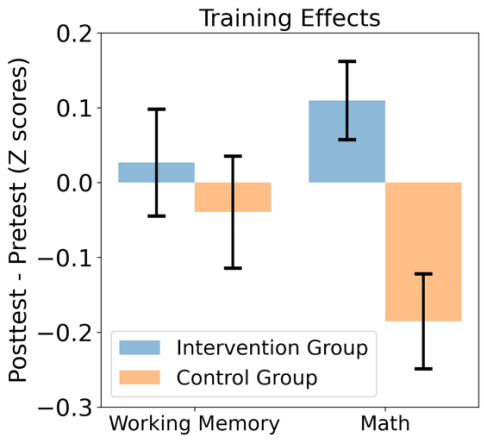
3.3 Results

Preliminary Results

To address research question 1, which aims to gain insights into the efficacy of synergistically training both executive function and math skills, we first conducted t-tests to compare the assessment performance changes between the intervention group and the control group. Specifically, we analyzed the gain scores, defined as the z score of the difference between posttest and pretest scores, for both working memory and math tasks separately as a function of group (details of the z score calculations were provided in the Methods section). The results showed no significant difference in working memory span changes from pretest to posttest between the intervention and control groups ($t = 0.62$, $p = .53$). However, the intervention group significantly improved in math performance as compared to the control group ($t = 3.59$, $p < .001$), see Figure 3.6.

Figure 3.6

Training Effects on Working Memory and Math Performance



Note. Error bars represent standard error of the means.

Next, focusing on the intervention group only, we calculated the correlations between participants' performance during the intervention and their working memory and mathematical growth as assessed before and after the training. Specifically, we utilized one key indicator from the training data: the improvement made during the training period (calculated as the difference between the maximum performance achieved and the performance in session 1). The correlation between training improvement and participants' working memory performance change was r = .11, p = .11. However, there was a significant positive correlation between training improvement and participants' math performance change, r = .16, p = .009.

Training Performance Prediction

To address research question 2, we focused on predicting each individual's training improvement by training machine learning models using performance data from the initial sessions as predictors. These models aimed to forecast the improvement made during the entire training period, defined as the difference between the maximum performance achieved over the entire intervention period and the performance in session 1. For example, if a participant achieved their maximum performance in session 7, their overall training improvement would be calculated as the difference between their performance in session 7 and session 1. Additionally, we compared the predictive performance of different models based on the number of sessions included as predictors. For instance, we compared models using data from the first two sessions with models using data from the first three or four sessions. Although the predictive power for participants' training improvements increased with the number of training sessions included as predictors, our goal was to achieve accurate predictions as early as possible, ideally with fewer initial sessions included as predictors.

To make all machine learning models based on the number of sessions included as predictors comparable, we ensured consistent samples and target values across models. Given that participants completed 7.08 sessions of Fluency+ games (SD = 3.03) during the three-week training period on average, we included only those participants who completed at least seven sessions in the predictive model (N=180). This allowed us to build and test models using performance data from the first session up to the first six sessions as predictors. Since six sessions represented half of the prescribed training program (i.e., participants were encouraged to complete twelve sessions over three weeks), predicting overall improvements before reaching this midpoint would be more efficient.

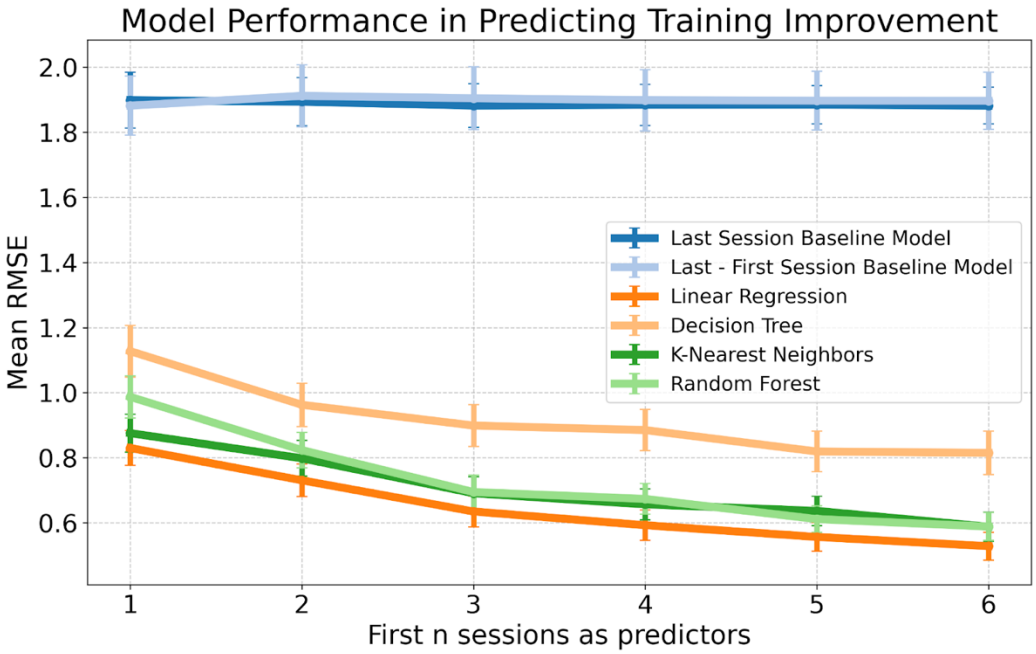
We applied several regression algorithms to train the models, including Linear Regression, Decision Tree, K-Nearest Neighbors, and Random Forest. Besides, we also included Last Session and Last - First Session estimation as our baseline models, which means using the last session performance (e.g. the second session performance if first two sessions performance as predictors) or the difference of last session minus first session performance as predicted training improvement. In particular, we used Root-mean-square deviation (RMSE) to evaluate model performance. The lower RMSE represents less differences between true value and predicted value which means higher predictive power. Given that our sample size was relatively small compared to typical sample sizes in the machine learning field, we adopted the Leave-One-Out Cross-Validation (LOOCV) approach in all algorithms (Vehtari, Gelman & Gabry, 2017). With this approach, we were able to split the dataset into a training set and a testing set, using all but one observation as part of the training set, and we repeated this process n times (where n is the total number of observations in the dataset), leaving out a different observation from the training set each time. We then calculated the average of all of the RMSEs in cross-validation sets.

Figure 3.7 presented the model performance for predicting participants' training improvements using the various algorithms implemented here. The results indicated that the Linear Regression algorithm exhibited the highest predictive power, significantly outperforming the two baseline models, as the RMSE was significantly lower than that of the baseline models. As expected, including more sessions' performance as predictors increased the accuracy of the predictions for training improvements. Using the performance data from the first six sessions as predictors resulted in the lowest RMSE value.

However, the research goal for research question 2 was early accurate prediction of training improvements. To do so, we conducted t-tests on the cross-validation RMSEs between the model using six sessions and models using fewer sessions. Notably, the prediction accuracy using data from the first three sessions did not differ significantly from using six sessions to predict training improvements ($t = 1.67$, $p = .10$). In contrast, using only the first two sessions as predictors resulted in a significantly higher RMSE compared to the six-session model ($p = .002$). Therefore, using the performance data from the first three sessions as predictors demonstrated the earliest acceptable predictive power. A good RMSE is often smaller than the standard deviation of the target variable, indicating good model performance if RMSE is significantly smaller (Willmott, 1982). Considering that the standard deviation of the target values was 1.22, the model using the first three sessions as predictors had an average RMSE of 0.63. This means the prediction error was within approximately 0.51 SD of the target variable, indicating good predictive power.

Figure 3.7

Models' Performance in Predicting Training Improvement



Note. Regression Model Performance by Different Algorithms in predicting training improvement. X-axis is the number of sessions used as predictors (first n sessions' performance as predictors); Y-axis represents the average RMSE across all cross-validation sets. Error bars represent standard error of the means.

Grouping Based on Early Learning Performance

Since the initial three sessions could predict overall training improvement, we chose these sessions as our source of early learning performance data to better understand participants' learning trajectories over the three-week training period. To answer research question 3, we employed K-means clustering to group students based on their performance during the first three sessions. The optimal number of clusters was determined to be two. Figure 3.8 illustrated that a two-cluster model provided a better fit compared to models

with other numbers of clusters. The clustering's effectiveness was evaluated using the Silhouette score and the Davies-Bouldin (DB) score, where a higher Silhouette score and a lower DB score indicate better clustering performance. The learning trajectories of these two groups were depicted in Figure 3.9. Two distinct patterns emerged: the orange group, which we named the Bamboo group (N=95), as they initially scored higher and demonstrated rapid improvement, akin to the fast growth characteristic of bamboo. Conversely, the blue group was referred to as the Aloe group (N=204), as they started with lower scores and showed slower improvement, similar to the gradual growth of aloe plants.

Figure 3.8

Clustering Goodness Evaluation

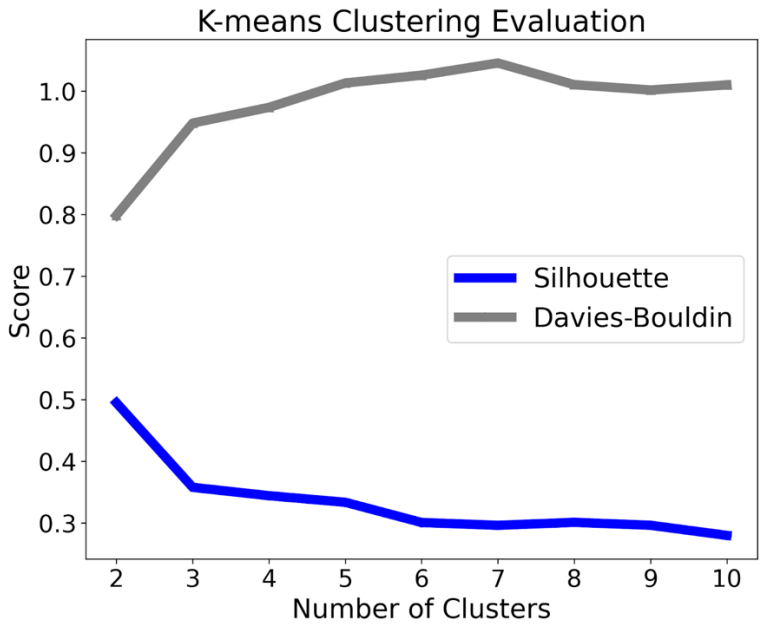
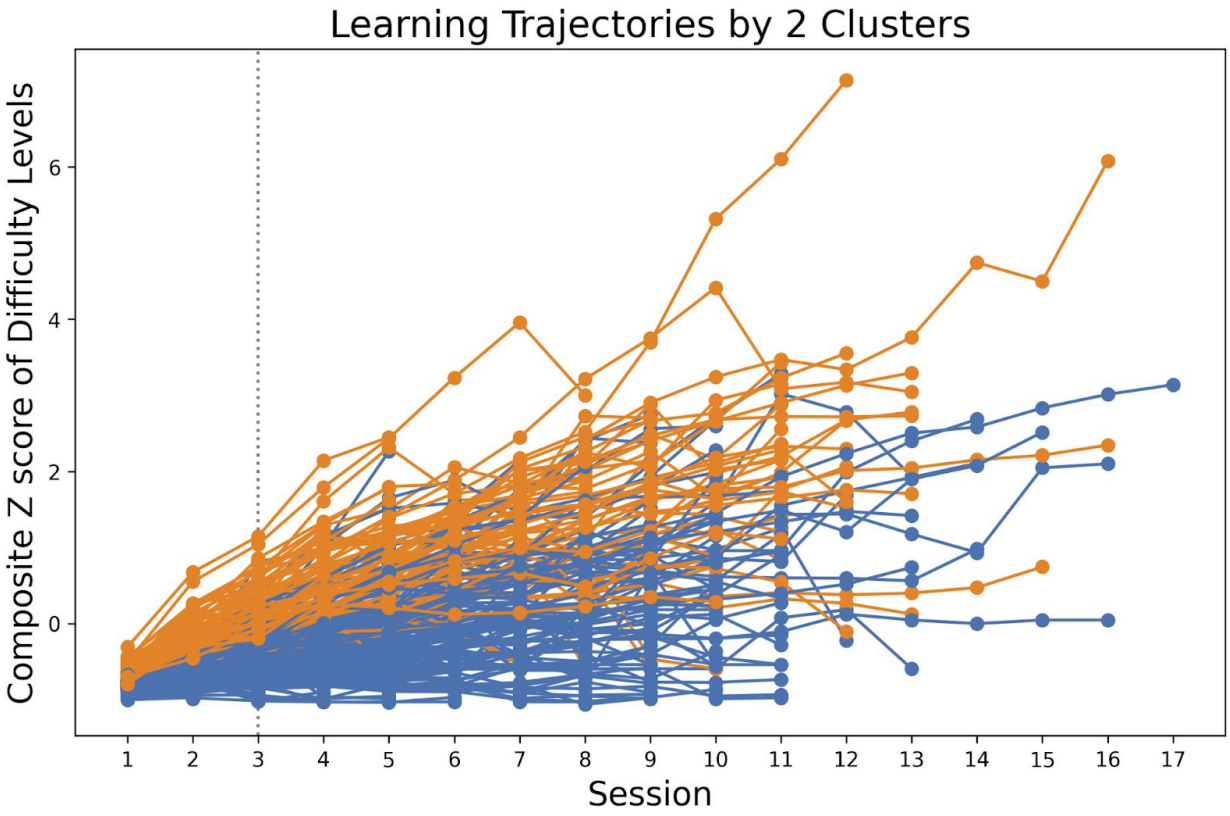


Figure 3.9

Clustering Results Based on First Three Sessions Performance



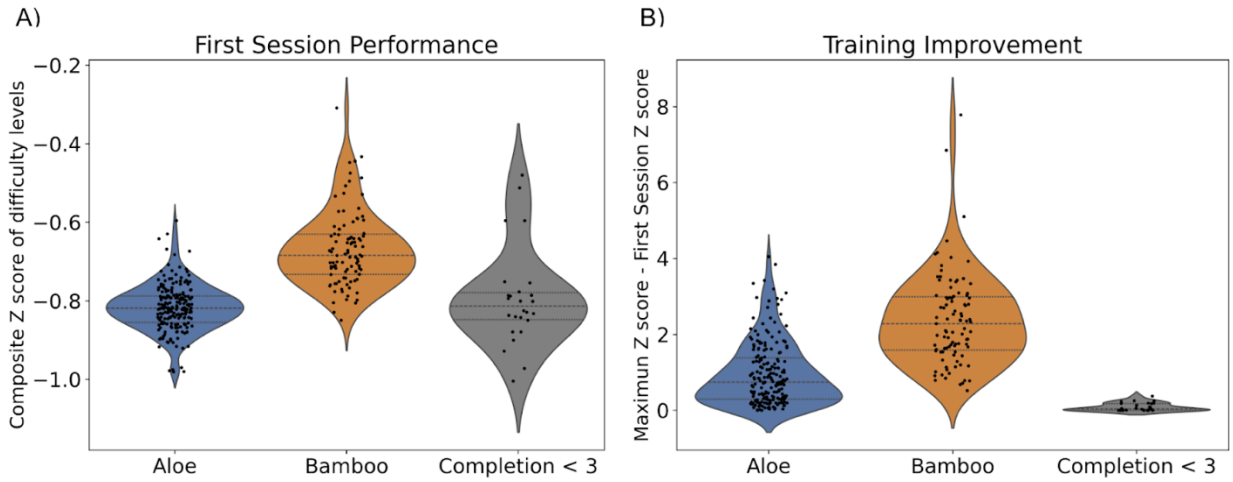
Note. Orange represented the "Bamboo" group; Blue represented the "Aloe" group.

A closer examination revealed that in Figure 3.10A, the Aloe group performed lower than the Bamboo group in the first session, with almost all members falling below the 25th percentile of the Bamboo group. The gray group (N=26) in the figure represented participants who were excluded from the clustering analysis due to only completing one or two sessions. In Figure 3.10B, the Bamboo group exhibited greater overall improvement than the Aloe group. However, some participants within the Aloe group showed similar

progress to those in the Bamboo group, as indicated by their placement above the -1 SD of the Bamboo group's performance.

Figure 3.10

Comparisons Between Aloe and Bamboo Groups



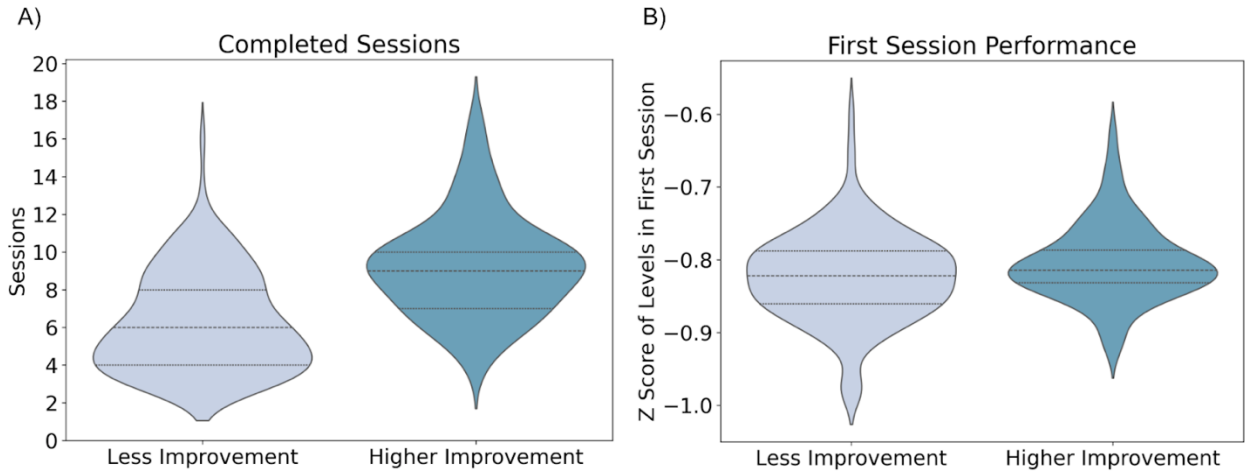
Detailed Examination of Aloe Group Subgroups

Further analysis within the Aloe group revealed differences between subgroups with higher and lower training improvements. The subgroup with higher improvement (above -1 SD of Bamboo group, N=64), comparable to the training improvement of Bamboo group, completed an average of nine sessions. In contrast, the lower improvement subgroup (N=140) completed only six sessions on average (Figure 3.11A). Further t-test results showed a significant difference between the two groups in the number of sessions completed ($t = 7.73$, $p < .001$). Though these two subgroups showed differences in their initial performance levels ($t = 2.16$, $p = .03$), their average values were in same range

compared to the Bamboo group (Less improvement group = -0.82, Higher improvement group = -0.80, Bamboo group = -0.67), as shown in Figure 3.11B.

Figure 3.11

Subgroup Comparisons Within the Aloe Group



Note. Violin plots compared less improvement and higher improvement subgroups within the Aloe group. The dashed lines within the violin plots represent the quartiles of the data.

3.4 Discussion

Our study conducted a novel intervention combining executive function with math tasks in three games for 4th and 5th graders over three weeks. The results showed that the intervention group improved their math assessment performance after training, but not their working memory performance compared to the control group. Analyzing only the intervention data, we found that using the performance from the first three sessions as predictors could forecast training improvement across the entire training period by applying a linear regression algorithm. This prediction result allowed us to identify

participants' learning patterns based on their first three sessions' performance. We found that participants could be clustered into two groups: a Bamboo group, which initially showed higher scores and greater improvement, and an Aloe group, which initially had lower scores and improved more slowly. Within the Aloe group, however, a higher improvement subgroup showed similar progress to the Bamboo group. This growth in the higher improvement subgroup could be attributed to the fact that they completed more training sessions compared to the subgroup who showed less improvement.

Our results were consistent with the findings in Chapter 2, where the combined intervention led to improvements in math performance but did not result in significant effects on working memory tasks. This indicates that the combined training approach is effective in enhancing math skills. In the intervention used in this study, we integrated executive function tasks into math tasks. This approach likely activated a wider range of brain regions during training (Deprez et al., 2013; Ward et al., 2019) and improved assessment performance by more effectively utilizing executive function in math tasks, reflecting real-world math problem-solving. Additionally, we found a correlation between improvement during training and gains in math assessments. This result suggested that the math training effect may have been partially driven by performance changes during the training period. This evidence underscored the importance of predicting training performance and highlighted that early prediction of training improvement could help identify students who might not show progress during training and in post-training assessments. Consistent with Chapter 2, we did not observe training effects on working memory skills. This may be due to several factors: 1) we did not have enough working memory training sessions compared to Zhang et al. (2019), which demonstrated effective

improvements in children's working memory with 20 sessions of dedicated training; 2) we only had one game primarily focused on improving working memory, leading to insufficient exposure to working memory components; 3) we had only one measure of working memory, which might not have been sensitive enough. Additionally, we did not conduct a follow-up test, where more pronounced effects in working memory training are typically observed (Jaeggi et al., 2011; Pugin et al., 2014).

In terms of individual differences, our findings aligned with previous studies on predicting cognitive training performance using initial performance data. For example, previous studies showed that baseline cognitive levels could predict training improvement (Guye et al.,2017; Kalbe et al., 2018; Ørskov et al., 2021). Our study extended this research by demonstrating that early learning performance could predict overall training improvement and by examining the optimal time point for accurate early predictions. Specifically, we found that training improvement could be predicted as early as the first three sessions with good predictive power. As the first exploratory study to test the predictive power of early learning performance using machine learning models, our research contributed to the growing body of literature supporting the use of these models for early detection of cognitive performance. For instance, studies by Delmastro et al. (2020) and Rykov et al. (2024) used wearable data and machine learning models to predict cognitive impairment. Our work also supported the strategic use of initial learning data to forecast and potentially optimize cognitive training trajectories, as seen in session-based recommendation studies where previous session performance predicted subsequent session performance to recommend appropriate training difficulty or content (Quadrana et al., 2017; Sandeep et al., 2020). Overall, our prediction results demonstrated that early

session data could provide valuable insights into potential training benefits, allowing for timely adjustments and additional assistance early in the intervention process.

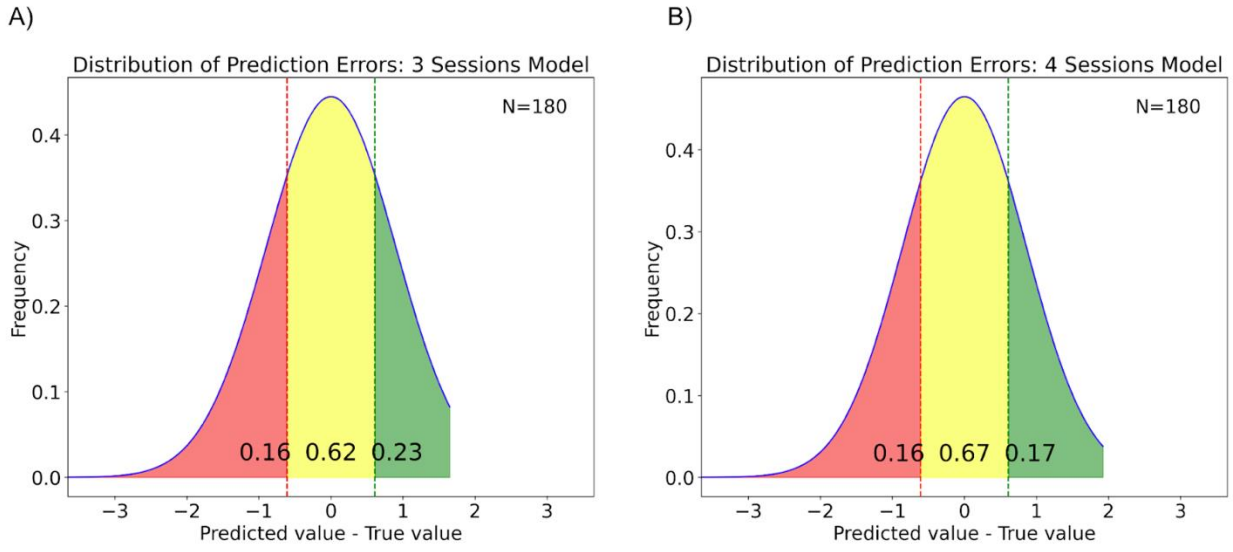
Critically, we found that some participants who started with lower performance but completed more training sessions achieved substantial benefits, essentially catching up to those with higher initial performance. Specifically, despite starting with lower scores, some participants in the Aloe group achieved improvements comparable to those in the Bamboo group by completing enough training sessions. This suggests that our interventions do not strictly follow a 'rich-get-richer' phenomenon; rather, individuals who initially struggled could still achieve significant learning gains similar to fast learners, provided they completed a sufficient number of training sessions. This highlights the importance of adhering to the intervention's regimen to maximize engagement and effectiveness (Chacko et al., 2014; Jones et al., 2020), especially considering that most students completed only half or less of the prescribed regimen. Several factors can affect students' commitment to an intervention regimen, such as a dense cognitive training schedule (Katz et al., 2018), teachers' attitudes and beliefs about a given educational protocol (Shawn Green et al., 2019), and entertainment during training (Jaeggi et al., 2014). For future implementation of the training program, we could consider using a more flexible training schedule, incorporating teachers' attitudes to motivate students, and adding more entertainment components to our training games to increase student engagement.

Several limitations should be acknowledged. First, despite using the leave-one-out method, the dataset had limited sample sizes (N=180) for those who completed seven sessions (half of the prescribed sessions), as many students failed to complete an adequate number of sessions. Second, the reliance on the first three sessions as the primary data

source introduced potential bias, as early performance may not always have been indicative of long-term outcomes for all participants. Third, there was the practical consideration of acceptable prediction error tolerance. Our analysis showed that the three-sessions model had 62% of cases within a "safe" error range, defined as ±0.5 standard deviations (SD) of the true values, while the four-sessions model improved this to 67% (see Figure 3.12). Determining an acceptable level of error for practical application remains an open question.

Figure 3.12

Prediction Error Distribution



Note. Prediction Error Distribution of A) using the first 3 sessions as predictors; B) using the first 4 sessions as predictors. Yellow represents the range where the distance between the predicted value and the true value is within ±0.5 standard deviation (SD) of the true value. Red indicates predicted values that are more than 0.5 SD above the true value, while green indicates predicted values that are more than 0.5 SD below the true value.

Another limitation was that we had to aggregate data across several datasets, leading us to rely on a reduced set of common measures. Specifically, our assessment measures for executive function were limited. A combination of tasks that could use a latent variable or composite measure would likely have been more effective. Additionally, our randomization was done at the classroom level for convenience, which can introduce bias and reduce the internal validity of the study. Furthermore, our control group could still have had contact with the intervention group, potentially influencing their performance and introducing contamination effects. These issues can be addressed in the follow-up study planned for summer 2024, where we aim to improve the study design by ensuring individual randomization, implementing a no-contact control group, and incorporating a broader range of assessment tasks to better measure executive function and math skills.

Overall, this study demonstrates that using the first three sessions as an early data source can effectively identify participants who may not benefit sufficiently from cognitive training. This identification point offers a critical timing for adjustments, such as gathering feedback or implementing motivational techniques. By examining the first three sessions, we identified two distinct learning patterns among participants. Although participants who initially performed better (Bamboo group) showed more improvement, some participants who started with lower performance but completed more training sessions (Aloe group) also achieved substantial benefits. This finding highlights the potential for significant improvement among initially lower-performing participants if they persist with more training sessions.

Future research should focus on expanding the dataset to include more diverse participant groups and exploring additional target variables beyond training improvement metrics. Investigating individual learning patterns, curves, or rates could yield more accurate forecasts of each participant's learning trajectory. This approach would facilitate personalized adjustments during training, optimizing the learning experience and tailoring it to the specific needs of each student. Moreover, enhancing the model's predictive power through advanced algorithms and more comprehensive data will facilitate personalized cognitive training programs, ultimately optimizing learning experiences and outcomes.

DISCUSSION

This dissertation explored the impact of individual differences on cognitive training and learning outcomes across three studies.

Study 1 focused on young adults' working memory training and used machine learning models to predict inter-individual variations in learning trajectories during training based on individual differences. Three learning patterns emerged as a result of clustering analyses: “Unicorns,” “Tortoises,” and “Hares,” named based on their early learning slopes and turning points. Unicorns demonstrated high initial performance and quick learning. Tortoises and hares showed lower initial performance, with gradual improvements, but differed as tortoises continued to improve longer, achieving higher final performance, while hares' performance leveled off earlier. In the next step, a binary tree model was developed to predict the three participant groups. Openness and baseline working memory capacity were the highest contributing factors in predicting Unicorns. For the Non-Unicorn group, a second model predicted whether they were hares or tortoises, showing that openness and video game background were the most important factors. This study demonstrated that learning trajectories can indeed be predicted and identified key factors, serving as a testbed to explore prediction models for forecasting learning patterns before training.

While Study 1 explored the prediction of learning trajectories based on individual difference factors, Study 2 focused on another major factor influencing learning outcomes: the intervention program itself, and how individual differences might interact with intervention efficacy. Specifically, Study 2 uncovered the potential differential effects of domain-specific and domain-general training games on children's math and working

memory learning outcomes. Building on the findings of Study 1 and previous research emphasizing the contribution of baseline skills to learning outcomes, Study 2 further explored the interaction between baseline skills and different types of interventions. The results of Study 2 suggested that an intervention incorporating both working memory and numerical knowledge training generally yielded the most substantial improvements in math learning. It is also noteworthy that for both math and working memory outcomes, individuals with lower initial math skills tended to benefit more from single-type (domain-specific) training, whereas those with higher initial skills benefited from combined training. The moderation effect of baseline skills further validated their role, corroborating with Study 1's finding that baseline skills play an important role in affecting learning outcomes.

Building on Study 2's findings, which suggested that combined working memory and math skills interventions generally yield larger training effects, Study 3 implemented a novel intervention that combined executive function training with math fluency training. 4th and 5th graders were asked to play these games over the course of three weeks. Given Study 1's findings on the predictive power of individuals' existing skills and characteristics, Study 3 further explored how initial training session performance could predict overall improvement.

Study 3 found that using performance data from the first three sessions as predictors could forecast training improvement across the entire period using a linear regression algorithm. This prediction allowed us to identify participants' learning patterns based on their initial performance. Participants were clustered into two groups: the Bamboo group, which showed higher initial scores and greater improvement, and the Aloe group, which started with lower scores and improved more slowly. Within the Aloe group,

a higher improvement subgroup showed progress similar to the Bamboo group, likely because they completed more training sessions than those with less improvement.

Combining the findings from the three studies provided several important insights. Firstly, both studies 1 and 3 in this dissertation found individual differences in learning, suggesting that people may have different learning rates. Although a recent publication by Koedinger et al. (2023) found that participants exhibited similar learning rates across different courses, this does not undermine the reliability of the individual differences observed in my research. The current study focused on specific tasks' cognitive learning process, while Koedinger et al.'s research examined learning rates across various types of courses, which may have averaged out individual variations. Additionally, they assumed all individuals followed a linear learning curve, which could have minimized individual differences by treating variations as error terms in the linear model. Therefore, despite their findings of similar learning rates, their results do not conflict with the individual differences in learning demonstrated in my study. Instead, these findings underscore the importance of developing individualized, non-linear models to more accurately represent the learning process. Secondly, in line with the rich-get-richer phenomenon, we observed that individuals with higher baseline skills or initial performance tended to show more improvement in cognitive training. However, this was not always the case. In Studies 1 and 3, participants with higher baseline skills or better initial performance demonstrated faster and greater learning improvement during training compared to those with lower initial skills. In contrast, although individuals with higher baseline skills performed better in posttests and follow-up tests in Study 2, those with lower baseline skills made more significant improvement in assessment tasks. Additionally, Study 3 showed that some

participants with lower baseline skills achieved substantial improvement, similar to those with higher baseline skills, likely due to completing more training sessions. These results suggested that when further studies consider the rich-get-richer phenomenon, it is important to distinguish between training improvement and transfer effects and to consider the completion rates.

The effectiveness of combined interventions was demonstrated in both Studies 2 and 3. Study 2 implemented a design where participants underwent five sessions of working memory training followed by five sessions of numerical knowledge training. This scaffolded approach showed a trend toward transfer effects in posttests and significant transfer effects in follow-up math performance after three months. Study 3 expanded this approach by integrating executive function and math tasks into a single training task. This method resulted in a significant transfer effect in math post-test performance, suggesting that combining these skills in one task might be more effective. It likely engaged broader brain regions during training (Deprez et al., 2013; Ward et al., 2019) and enhanced assessment performance by better applying executive function in math tasks, mirroring real-world math task processing.

Interestingly, neither Study 2 nor Study 3 found a transfer effect on working memory skills. One possible explanation is that the design of the working memory training tasks in both studies may not have effectively induced a training effect. Study 2 combined data from three interventions, with the training trajectory from the third data collection reported in DePascale et al., 2024. In this report, the accuracy of the working memory training decreased over the five sessions (in their Figure 5), and the working memory span set size did not show significant change. This indicated that participants' performance did

not improve over the course of the training. In Study 3, training performance was represented by a composite z-score of the progress in three game levels, with each game's level determined by both one executive function component and math performance. Therefore, the progress in Study 3's training performance did not necessarily reflect improvements in the working memory component. Going forward, we will break down the training data from Studies 2 and 3 to focus specifically on the progress in working memory training. This analysis may provide more insight into whether the design of the working memory training tasks was ineffective. Another explanation could be due to the duration of training, which was considerably shorter compared to previous studies that implemented 20 sessions (Zhang et al., 2019).

Furthermore, both Studies 1 and 3 explored the predictive power of pre-intervention factors and early performance during intervention on training improvement. The combined results indicated that initial performance played a significant role in predicting learning improvement. This suggested that pre-existing performance or early learning performance could be used to predict training outcomes, allowing for the selection of appropriate training programs or early adjustments to interventions. However, since Studies 1 and 3 focused only on executive function and/or math training, further research is needed to determine if the predictive power of initial performance generalizes to other cognitive training domains. Overall, these insights highlighted the importance of considering both initial abilities and the specific design, along with the duration of interventions when aiming to optimize cognitive training outcomes.

The implications of these findings are profound for both educational practice and cognitive training. The ability to predict learning trajectories and outcomes based on

individual differences facilitates the design of more effective and personalized training programs. This approach aligns with the goals of precision education, where interventions are tailored to meet the unique needs of each learner. The evidence supporting the interaction between learner attributes and intervention types suggests that educational programs should be flexible and adaptable. Implementing differentiated instruction strategies can cater to diverse student populations and enhance overall learning outcomes.

Furthermore, using predictive models to monitor and adjust training programs in real-time could significantly enhance educational strategies. Early identification of at-risk students allows for timely interventions, reducing the likelihood of students falling behind and ensuring that all learners have the opportunity to succeed. Additionally, the study highlights the importance of participant adherence to training regimens. Factors such as the training schedule, motivational support, and the integration of engaging elements can influence commitment and effectiveness. Future programs should consider these factors to enhance student engagement and training outcomes.

In conclusion, this dissertation advances our understanding of personalized education by demonstrating the critical role of individual differences in cognitive training and learning outcomes. The findings collectively highlight the necessity of tailored interventions that consider personality traits, baseline abilities, and other individual factors. By leveraging machine learning and predictive models, we have the potential to design more effective training programs and timely adjustment that accommodate the diverse needs of learners, ultimately enhancing educational practices and outcomes. This dissertation paves the way for future studies to develop even more refined models and

interventions, contributing to the field of personalized education and promoting equitable learning opportunities for all students.

REFERENCES

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological bulletin*, 102(1), 3.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of experimental psychology: General*, 117(3), 288.

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: dynamics of ability determinants. *Journal of applied psychology*, 77(5), 598.

Ackerman, P. L., Bowen, K. R., Beier, M. E., & Kanfer, R. (2001). Determinants of individual differences and gender differences in knowledge. *Journal of educational psychology*, 93(4), 797.

Ackerman, P. L. (2007). New developments in understanding skilled performance. Current Directions in *Psychological Science*, 16(5), 235-239.

Albon, C. (2018). Machine learning with python cookbook: *Practical solutions from preprocessing to deep learning*. " O'Reilly Media, Inc.".

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of experimental child psychology*, *106*(1), 20-29.

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., ... & Durrleman, S. (2021). Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Medical Image Analysis*, 67, 101848.

Aschwanden, D., Aichele, S., Ghisletta, P., Terracciano, A., Kliegel, M., Sutin, A. R., ... & Allemand, M. (2020). Predicting cognitive impairment and dementia: A machine learning approach. *Journal of Alzheimer's Disease*, *75*(3), 717-728.

Au, J., Katz, B., Buschkuehl, M., Bunarjo, K., Senger, T., Zabel, C., ... & Jonides, J. (2016). Enhancing working memory training with transcranial direct current stimulation. *Journal of cognitive neuroscience*, 28(9), 1419-1432.

Au, J., Buschkuehl, M., Duncan, G. J., & Jaeggi, S. M. (2016). There is no convincing evidence that working memory training is NOT effective: A reply to Melby-Lervåg and Hulme (2015). *Psychonomic Bulletin & Review*, 23(1), 331-337.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of social and clinical psychology*, 4(3), 359-373.

Bastian, C. C. von, & Oberauer, K. (2014). Effects and mechanisms of working memory training: A review. *Psychological Research*, 78(6), 803–820. https://doi.org/10.1007/s00426-013-0524-6

Bekoff, M. (1988). Motor training and physical fitness: Possible short-and long-term influences on the development of individual differences in behavior. Developmental Psychobiology: The Journal of the International Society for *Developmental Psychobiology*, 21(6), 601-612.

Berard, A. V., Cain, M. S., Watanabe, T., & Sasaki, Y. (2015). Frequent Video Game Players Resist Perceptual Interference. *PLOS ONE*, 10(3), e0120011. https://doi.org/10.1371/journal.pone.0120011

Benavides-Varela, S., Callegher, C. Z., Fagiolini, B., Leo, I., Altoè, G., & Lucangeli, D. (2020). Effectiveness of digital-based interventions for children with mathematical learning difficulties: A meta-analysis. *Computers & Education*, *157*, 103953.

Bergman-Nutley, S., & Klingberg, T. (2014). Effect of working memory training on working memory, arithmetic and following instructions. *Psychological research*, 78(6), 869-877.

Berrar, D. (2018). Cross-Validation. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Blair, C., & Raver, C. C. (2014). Closing the achievement gap through modification of neurocognitive and neuroendocrine function: Results from a cluster randomized controlled trial of an innovative approach to the education of children in kindergarten. *PloS one*, 9(11), e112393.

Blotnicky, K. A., Franz-Odendaal, T., French, F., & Joy, P. (2018). A study of the correlation between STEM career knowledge, mathematics self-efficacy, career interests, and career activities on the likelihood of pursuing a STEM career among middle school students. *International journal of STEM education*, 5(1), 1-15.

Boger-Mehall, S. R. (1996). Cognitive flexibility theory: Implications for teaching and teacher education. *In Society for Information Technology & Teacher Education International Conference* (pp. 991-993). Association for the Advancement of Computing in Education (AACE).

Bogg, T., & Lasecki, L. (2015). Reliable gains? Evidence for substantially underpowered designs in studies of working memory training transfer to fluid intelligence. *Frontiers in psychology*, 5, 1589.

Borella, E., Carbone, E., Pastore, M., De Beni, R., & Carretti, B. (2017). Working memory training for healthy older adults: the role of individual characteristics in explaining short-and long-term gains. *Frontiers in human neuroscience*, 11, 99.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, 21(1), 1–16. https://doi.org/10.1111/j.2044-8260.1982.tb01421.x

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental neuropsychology*, 33(3), 205-228.

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, 8(1), 36-41.

Bürki, C. N., Ludwig, C., Chicherio, C., & Ribaupierre, A. de. (2014). Individual differences in cognitive plasticity: An investigation of training curves in younger and older adults. *Psychological Research*, 78(6), 821–835. https://doi.org/10.1007/s00426-014-0559-3

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of experimental child psychology*, 33(3), 386-404.

Cassilhas, R. C., Tufik, S., & de Mello, M. T. (2016). Physical exercise, neuroplasticity, spatial learning and memory. *Cellular and Molecular Life Sciences*, 73(5), 975-983.

Chacko, A., H Newcorn, J., Feirsen, N., & Z Uderman, J. (2010). Improving medication adherence in chronic pediatric health conditions: a focus on ADHD in youth. *Current pharmaceutical design*, *16*(22), 2416-2423.

Chamorro-Premuzic, T., & Furnham, A. (2009). Mainly Openness: The relationship between the Big Five personality traits and learning approaches. *Learning and individual Differences*, 19(4), 524-529.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

Chodura, S., Kuhn, J. T., & Holling, H. (2015). Interventions for children with mathematical difficulties. *Zeitschrift für Psychologie*.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American educational research journal*, 45(2), 443-494.

Clements, D. H., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, *36*, 79-90.

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain.*

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory?. *Progress in brain research*, 169, 323-338.

Cowan, R., & Powell, D. (2014). The contributions of domain-general and numerical factors to third-grade arithmetic skills and mathematical learning disability. *Journal of educational psychology*, 106(1), 214.

Cronbach, L. J., & Snow, R. E. (1969). Individual Differences in Learning Ability as a Function of Instructional Variables. *Final Report*.

Cronbach, L. J., and Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Diamond, A., & Ling, D. S. (2020). Review of the evidence on, and fundamental questions about, efforts to improve executive functions, including working memory. *Cognitive and working memory training: Perspectives from psychology, neuroscience, and human development*, 143-431.

Datu, J. A. D. (2021). Beyond passion and perseverance: Review and future research initiatives on the science of grit. *Frontiers in Psychology*, 11, 545526.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44(11), 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

Delmastro, F., Di Martino, F., & Dolciotti, C. (2020). Cognitive training and stress detection in mci frail older people through wearable sensors and machine learning. *Ieee Access*, 8, 65573-65590.

Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, 115(8), 1690-1692.

DePascale, M., Feng, Y., Lin, G. C., Barkin, R., Akhavein, K., Tavassolie, N., ... & Jaeggi, S. M. (2024). Uncovering the reciprocal relationship between domain-specific and domain-general skills: Combined numerical and working memory training improves children's mathematical knowledge. *Contemporary Educational Psychology*, 76, 102252.

Deprez, S., Vandenbulcke, M., Peeters, R., Emsell, L., Amant, F., & Sunaert, S. (2013). The functional neuroanatomy of multitasking: combining dual tasking with a short term memory task. *Neuropsychologia*, 51(11), 2251-2260.

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: effects of feedback and training. *Frontiers in human neuroscience*, 6, 68.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, 64(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2(1), 1-17.

Dong, Y., Clements, D. H., Sarama, J., Dumas, D., Banse, H. W., & Day-Hess, C. (2022). Mathematics and executive function competencies in the context of interventions: A quantile regression analysis. The Journal of Experimental Education, 90(2), 297-318.

Duckworth, Angela L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology,* 92(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, Angela Lee, & Quinn, P. D. (2009). Development and Validation of the Short Grit Scale (Grit–S). *Journal of Personality Assessment,* 91(2), 166–174. https://doi.org/10.1080/00223890802634290

Duncan, G. J., & Magnuson, K. (2012). Socioeconomic status and cognitive functioning: moving from correlation to causation. *Wiley Interdisciplinary Reviews: Cognitive Science,* 3(3), 377-386.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.

Eason, S. H., Scalise, N. R., Berkowitz, T., Ramani, G. B., & Levine, S. C. (2022). Widening the lens of family math engagement: A conceptual framework and systematic review. *Developmental Review*, *66*, 101046.

Farr, B. J. (1971). Individual differences in learning: Predicting one's more effective learning modality. *The Catholic University of America.*

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of experimental child psychology*, 123, 53-72.

Fellman, D., Jylkkä, J., Waris, O., Soveri, A., Ritakallio, L., Haga, S., ... & Laine, M. (2020). The role of strategy use in working memory training outcomes. *Journal of Memory and Language*, 110, 104064.

Feng, Y., Buschkuehl, M., Grose, G., Jaeggi, S.M., Ramani, G., Eunice G., DePascale, M. (2022 June). Testing the benefits of an embedded math fluency and executive function intervention in children. *Poster presented at the 4th MCLS 2022 Annual Meeting*, Antwerp, Belgium.

Feng, Y., Pahor, A., Seitz, A. R., Barbour, D. L., & Jaeggi, S. M. (2023). Unicorn, Hare, or Tortoise? Using Machine Learning to Predict Working Memory Training Performance. *Journal of Cognition*, *6*(1).

Fleishman, E. A. (1967). Individual differences and motor learning. *American Institutes for Research.*

Flak, M. M., Hol, H. R., Hernes, S. S., Chang, L., Engvig, A., Bjuland, K. J., ... & Løhaugen, G. C. (2019). Adaptive computerized working memory training in patients with mild cognitive impairment. A randomized double-blind active controlled trial. *Frontiers in psychology,* 10, 807.

Forbes, L., Gutierrez, D., & Johnson, S. K. (2018). Investigating adherence to an online introductory mindfulness program. *Mindfulness*, *9*, 271-282.

Foster, J. L., Harrison, T. L., Hicks, K. L., Draheim, C., Redick, T. S., & Engle, R. W. (2017). Do the effects of working memory training depend on baseline ability level? *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 43(11), 1677. https://doi.org/10.1037/xlm0000426

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., ... & Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical versus general cognitive abilities?. *Developmental psychology*, *46*(6), 1731.

Fuhs, M. W., Nesbitt, K. T., Farran, D. C., & Dong, N. (2014). Longitudinal associations between executive functioning and academic skills across content areas. *Developmental psychology*, *50*(6), 1698.

Fuhs, M. W., Hornburg, C. B., & McNeil, N. M. (2016). Specific early number skills mediate the association between executive functioning skills and mathematics achievement. *Developmental Psychology*, *52*(8), 1217.

Fuchs, L., Fuchs, D., Seethaler, P. M., & Barnes, M. A. (2020). Addressing the role of working memory in mathematical word-problem solving when designing intervention for struggling learners. *ZDM*, 52, 87-96.

Gathercole, S. E., Brown, L., & Pickering, S. J. (2003). Working memory assessments at school entry as longitudinal predictors of National Curriculum attainment levels. *Educational and Child Psychology*, 20(3), 109-122.

Gathercole, S. E., Durling, E., Evans, M., Jeffcock, S., & Stone, S. (2008). Working memory abilities and children's performance in laboratory analogues of classroom activities. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition,* 22(8), 1019-1037.

Geary, D. C., & Hoard, M. K. (2005). Learning Disabilities in Arithmetic and Mathematics: Theoretical and Empirical Perspectives. In J. I. D. Campbell (Ed.), The Handbook of Mathematical Cognition (0 ed., pp. 271–286). Psychology Press. https://doi.org/10.4324/9780203998045-24

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: a 5-year longitudinal study. *Developmental psychology*, *47*(6), 1539.

Geary, D. C. (2013). Early foundations for mathematics learning and their relations to learning disabilities. *Current directions in psychological science*, 22(1), 23-27.

Geary, D. C., Nicholas, A., Li, Y., & Sun, J. (2017). Developmental change in the influence of domain-general abilities and domain-specific knowledge on mathematics achievement: An eight-year longitudinal study. *Journal of Educational Psychology*, 109(5), 680.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. *O'Reilly Media*.

Gliksman, Y., Berebbi, S., & Henik, A. (2022). Math Fluency during Primary School. *Brain Sciences*, 12(3), 371. https://doi.org/10.3390/brainsci12030371

Goodman, E., Adler, N. E., Kawachi, I., Frazier, A. L., Huang, B., & Colditz, G. A. (2001). Adolescents' Perceptions of Social Status: Development and Evaluation of a New Indicator. *PEDIATRICS*, 108(2), e31–e31. https://doi.org/10.1542/peds.108.2.e31

Green, C. S., & Newcombe, N. S. (2020). Cognitive training: How evidence, controversies, and challenges inform education policy. *Policy Insights from the Behavioral and Brain Sciences*, *7*(1), 80-86.

Grueso, S., & Viejo-Sobera, R. (2021). Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. *Alzheimer's research & therapy*, *13*, 1-29.

Guye, S., De Simoni, C., & von Bastian, C. C. (2017). Do Individual Differences Predict Change in Cognitive Training Performance? A Latent Growth Curve Modeling Approach. *Journal of Cognitive Enhancement,* 1(4), 374–393. https://doi.org/10.1007/s41465-017-0049-9

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.

Hambrick, D. Z., & Engle, R. W. (2003). The role of working memory in problem solving. *The psychology of problem solving*, 176-206.

Harrell, E. R., Roque, N. A., Boot, W. R., & Charness, N. (2021). Investigating message framing to improve adherence to technology-based cognitive interventions. *Psychology and aging*, 36(8), 974.

Hassler Hallstedt, M., Klingberg, T., & Ghaderi, A. (2018). Short and long-term effects of a mathematics tablet intervention for low performing second graders. *Journal of Educational Psychology*, *110*(8), 1127.

Hawes, Z., Nosworthy, N., Archibald, L., & Ansari, D. (2019). Kindergarten children's symbolic number comparison skills relates to 1st grade mathematics achievement: Evidence from a two-minute paper-and-pencil test. *Learning and Instruction*, 59, 21-33.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental science*, 12(4), F9-F15.

Hubert, M., & Veeken, S. V. der. (2008). Outlier detection for skewed data. *Journal of Chemometrics*, 22(3-4), 235–246. https://doi.org/10.1002/cem.1123

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, 105(19), 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, 108(25), 10081–10086. https://doi.org/10.1073/pnas.1103228108

Jaeggi, Susanne M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition,* 42(3), 464–480. https://doi.org/10.3758/s13421-013-0364-z

Jensen, A. R. (1967). How much can we boost IQ and scholastic achievement?

Jones, M. R., Katz, B., Buschkuehl, M., Jaeggi, S. M., & Shah, P. (2020). Exploring n-back cognitive training for children with ADHD. *Journal of attention disorders*, *24*(5), 704-719.

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of educational psychology*, 104(3), 647.

Kalbe, E., Bintener, C., Ophey, A., Reuter, C., Göbel, S., Klöters, S., ... & Kessler, J. (2018). Computerized cognitive training in healthy older adults: Baseline cognitive level and subjective cognitive concerns predict training outcome. *Health*, 10(01), 20.

Karbach, J., & Unger, K. (2014). Executive control training from middle childhood to adolescence. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.00390

Karbach, J., Strobach, T., & Schubert, T. (2015). Adaptive working-memory training benefits reading, but not mathematics in middle childhood. *Child Neuropsychology*, 21(3), 285-301.

Karbach, J., Könen, T., & Spengler, M. (2017). Who Benefits the Most? Individual Differences in the Transfer of Executive Control Training Across the Lifespan. *Journal of Cognitive Enhancement*, 1(4), 394–405. https://doi.org/10.1007/s41465-017-0054-z

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near-and far-transfer effects among children's executive function skills. *Psychological Bulletin*, *145*(2), 165.

Katz, B., Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2018). The effect of monetary compensation on cognitive training outcomes. *Learning and Motivation*, *63*, 77-90.

Katz, B., Jones, M. R., Shah, P., Buschkuehl, M., & Jaeggi, S. M. (2021). Individual Differences in Cognitive Training Research. In T. Strobach & J. Karbach (Eds.), Cognitive

Training: An Overview of Features and Applications (pp. 107–123). *Springer International Publishing*. https://doi.org/10.1007/978-3-030-39292-5_8

Kekäläinen, T., Terracciano, A., Tirkkonen, A., Savikangas, T., Hänninen, T., Neely, A. S., ... & Kokko, K. (2023). Does personality moderate the efficacy of physical and cognitive training interventions? A 12-month randomized controlled trial in older adults. *Personality and Individual Differences*, 202, 111957.

Kessels, R. P., van Zandvoort, M. J., Postma, A., Kappelle, L. J., & de Haan, E. H. (2000). The Corsi Block-Tapping Task: Standardization and normative data. *Applied Neuropsychology*, 7(4), 252–258. https://doi.org/10.1207/S15324826AN0704_8

Khan, N. A., & Hillman, C. H. (2014). The relation of childhood physical activity and aerobic fitness to brain function and cognition: a review. *Pediatric exercise science*, 26(2), 138-146.

Klingberg, T. (2010). Training and plasticity of working memory. Trends in Cognitive Sciences, 14(7), 317–324. https://doi.org/10.1016/j.tics.2010.05.002

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences*, *120*(13), e2221311120.

Könen, T., & Karbach, J. (2015). The benefits of looking at intraindividual dynamics in cognitive training data. *Frontiers in Psychology*, 6. https://doi.org/10.3389/fpsyg.2015.00615

Kroesbergen, E. H., Van't Noordende, J. E., & Kolkman, M. E. (2012). Number sense in low-performing kindergarten children: Effects of a working memory and an early math

training. *In Reading, writing, mathematics and the developing brain: Listening to many voices* (pp. 295-313). Dordrecht: Springer Netherlands.

Kroesbergen, E. H., van't Noordende, J. E., & Kolkman, M. E. (2014). Training working memory in kindergarten children: Effects on working memory and early numeracy. *Child Neuropsychology,* 20(1), 23-37.

Lachheb, A., Abramenka-Lachheb, V., & Huber, L. (2021). Challenges and opportunities in adhering to UDL principles to design online courses. *The Journal of Applied Instructional Design*, *10*(1), 97-123.

Laine, M., Fellman, D., Waris, O., & Nyman, T. J. (2018). The early effects of external and internal strategies on working memory updating training. *Scientific Reports*, 8(1), 4045. https://doi.org/10.1038/s41598-018-22396-5

Lee, K. (2023). Using working memory intervention to improve math performance: ill-conceived, poorly executed, or just not quite there yet?(Intervenciones en memoria de trabajo para mejorar el rendimiento matemático:¿ mal diseñadas, pobremente ejecutadas, o poco desarrolladas?). *Journal for the Study of Education and Development*, 46(1), 39-70.

Lee, J., Lee, H. J., Song, J., & Bong, M. (2021). Enhancing children's math motivation with a joint intervention on mindset and gender stereotypes. *Learning and Instruction*, *73*, 101416.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2), 451-461.

Lövdén, M., Brehmer, Y., Li, S. C., & Lindenberger, U. (2012). Training-induced compensation versus magnification of individual differences in memory performance. *Frontiers in human neuroscience*, 6, 141.

Luan, H., & Tsai, C. C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1), 250-266.

Lumsden, J., Edwards, E. A., Lawrence, N. S., Coyle, D., & Munafò, M. R. (2016). Gamification of cognitive assessment and cognitive training: a systematic review of applications and efficacy. *JMIR serious games*, *4*(2), e5888.

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. https://arxiv.org/abs/1705.07874v2

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1–6. *Developmental science*, 17(5), 714-726.

Maćkiewicz, M., & Cieciuch, J. (2016). Pictorial personality traits questionnaire for children (PPTQ-C)—A new measure of children's personality traits. *Frontiers in psychology*, 7, 498.

Martin, R. B., Cirino, P. T., Sharp, C., & Barnes, M. (2014). Number and counting skills in kindergarten as predictors of grade 1 mathematical skills. *Learning and individual differences*, 34, 12-23.

Matysiak, O., Kroemeke, A., & Brzezicka, A. (2019). Working memory capacity as a predictor of cognitive training efficacy in the elderly population. *Frontiers in Aging Neuroscience*, 11, 126.

McClelland, M. M., Tominey, S. L., Schmitt, S. A., Hatfield, B. E., Purpura, D. J., Gonzales, C. R., & Tracy, A. N. (2019). Red light, purple light! Results of an intervention to promote school readiness for children from low-income backgrounds. *Frontiers in Psychology*, *10*, 470058.

McCrae, R. R., & Sutin, A. R. (2009). Openness to experience. In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (pp. 257–273). The Guilford Press.

McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of experimental psychology: general*, 141(2), 302.

Meiran, N., Dreisbach, G., & von Bastian, C. C. (2019). Mechanisms of working memory training: Insights from individual differences. *Intelligence*, 73, 78–87. https://doi.org/10.1016/j.intell.2019.01.010

Miyake, A., & Shah, P. (Eds.). (1999). Models of working memory: Mechanisms of active maintenance and executive control. Cambridge University Press. https://doi.org/10.1017/CBO9781139174909

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental psychology*, *49*(2), 270.

Melby-Lervåg, M., & Hulme, C. (2016). There is no convincing evidence that working memory training is effective: A reply to Au et al.(2014) and Karbach and Verhaeghen (2014). *Psychonomic Bulletin & Review*, 23(1), 324-330.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of Machine Learning, second edition. *MIT Press.*

Munez, D., Lee, K., Bull, R., Khng, K. H., Cheam, F., & Rahim, R. A. (2022). Working memory and numeracy training for children with math learning difficulties: Evidence from a large-scale implementation in the classroom. *Journal of Educational Psychology*, 114, 1866. https://doi.org/10.1037/edu0000732

Nelson, G., & McMaster, K. L. (2019). The effects of early numeracy interventions for students in preschool and early elementary: A meta-analysis. *Journal of Educational Psychology*, *111*(6), 1001.

Nemmi, F., Helander, E., Helenius, O., Almeida, R., Hassler, M., Räsänen, P., & Klingberg, T. (2016). Behavior and neuroimaging at baseline predict individual response to combined mathematical and working memory training in children. *Developmental Cognitive Neuroscience*, 20, 43-51.

Nemmi, F., Nymberg, C., Helander, E., & Klingberg, T. (2016). Grit Is Associated with Structure of Nucleus Accumbens and Gains in Cognitive Training. *Journal of Cognitive Neuroscience*, 28(11), 1688–1699. https://doi.org/10.1162/jocn_a_01031

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine Learning in Psychometrics and Psychological Research. *Frontiers in Psychology,* 10, 2970. https://doi.org/10.3389/fpsyg.2019.02970

Ørskov, P. T., Norup, A., Beatty, E. L., & Jaeggi, S. M. (2021). Exploring individual differences as predictors of performance change during dual-n-back training. *Journal of Cognitive Enhancement*, 5(4), 480-498.

Pahor, A., Stavropoulos, T., Jaeggi, S. M., & Seitz, A. R. (2019). Validation of a matrix reasoning task for mobile devices. *Behavior Research Methods,* 51(5), 2256–2267. https://doi.org/10.3758/s13428-018-1152-2

Pahor, A., Collins, C., Smith-Peirce, R. N., Moon, A., Stavropoulos, T., Silva, I., Peng, E., Jaeggi, S. M.,

& Seitz, A. R. (2021). Multisensory Facilitation of Working Memory Training. *Journal of Cognitive Enhancement*, 5(3), 386–395. DOI: https://doi.org/10.1007/s41465-020-00196-y

Pahor, A., Seitz, A. R., & Jaeggi, S. M. (2022). Near transfer to an unrelated N-back task mediates the effect of N-back working memory training on matrix reasoning. *Nature Human Behaviour*, 6(9), 1243-1256.

Pahor, A., Mester, R. E., Carrillo, A. A., Ghil, E., Reimer, J. F., Jaeggi, S. M., & Seitz, A. R. (2022). UCancellation: A new mobile measure of selective attention and concentration. *Behavior Research Methods*.

Park, S., & Park, H. J. (2016). The effects of maternal parenting stress trajectory on affective parenting and socioemotional development of preschooler: Differences between employed and unemployed mothers. *Korean Journal of Child Studies,* 37(5), 33-52.

Park, J., Bermudez, V., Roberts, R. C., & Brannon, E. M. (2016). Non-symbolic approximate arithmetic training improves math performance in preschoolers. *Journal of experimental child psychology*, 152, 278-293.

Passolunghi, M. C., Mammarella, I. C., & Altoè, G. (2008). Cognitive abilities as precursors of the early acquisition of mathematical skills during first through second grades. *Developmental neuropsychology*, *33*(3), 229-250.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, 12(2011): 2825–2830.

Petscher, Y., Al Otaiba, S., & Wanzek, J. (2021). Study of the Factor Structure, Profiles, and Concurrent Validity of the Mindset Assessment Profile Tool for Elementary Students. *Journal of Psychoeducational Assessment*, 39(1), 74–88. https://doi.org/10.1177/0734282920943456

Prager, E. O., Ernst, J. R., Mazzocco, M. M., & Carlson, S. M. (2023). Executive function and mathematics in preschool children: Training and transfer effects. *Journal of Experimental Child Psychology*, 232, 105663

Pugin, F., Metz, A. J., Stauffer, M., Wolf, M., Jenni, O. G., & Huber, R. (2014). Working memory training shows immediate and long-term effects on cognitive performance in children. *F1000Research*, *3*.

Quadrana, M., Karatzoglou, A., Hidasi, B., & Cremonesi, P. (2017, August). Personalizing session-based recommendations with hierarchical recurrent neural networks. *In proceedings of the Eleventh ACM Conference on Recommender Systems* (pp. 130-137).

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and individual differences,* 20(2), 110-122.

Ramani, G. B., Jaeggi, S. M., Daubert, E. N., & Buschkuehl, M. (2017). Domain-specific and domain-general training to improve kindergarten children's mathematics. *Journal of Numerical Cognition*, 3(2), 468-495.

Ramani, G. B., Daubert, E. N., Lin, G. C., Kamarsu, S., Wodzinski, A., & Jaeggi, S. M. (2020). Racing dragons and remembering aliens: Benefits of playing number and working memory games on kindergartners' numerical knowledge. *Developmental Science*, 23(4). https://doi.org/10.1111/desc.12908

Rebok, G. W., Carlson, M. C., & Langbaum, J. B. (2007). Training and maintaining memory abilities in healthy older adults: traditional and novel approaches. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 62(Special_Issue_1), 53-61.

Redick, T. S. (2019). The Hype Cycle of Working Memory Training. *Current Directions in Psychological Science*, 28(5), 423–429. https://doi.org/10.1177/0963721419848668

Reynvoet, B., Vanbecelaere, S., Depaepe, F., & Sasanguie, D. (2021). Intervention studies in math: A metareview. In *Heterogeneous contributions to numerical cognition* (pp. 283-308). Academic Press.

Ribner, A. D. (2020). Executive function facilitates learning from math instruction in kindergarten: Evidence from the ECLS-K. *Learning and instruction*, 65, 101251.

Ribner, A. D., & Libertus, M. (2021). Mechanisms Underlying Transfer from Domain-Specific and Domain-General Cognitive Training to Children's Math Skills.

Robert, C., Borella, E., Fagot, D., Lecerf, T., & De Ribaupierre, A. (2009). Working memory and inhibitory control across the life span: Intrusion errors in the Reading Span Test. *Memory & cognition*, 37(3), 336-345.

Rykov, Y. G., Patterson, M. D., Gangwar, B. A., Jabar, S. B., Leonardo, J., Ng, K. P., & Kandiah, N. (2024). Predicting cognitive scores from wearable-based digital physiological features using machine learning: data from a clinical trial in mild cognitive impairment. *BMC medicine*, *22*(1), 36.

Sánchez-Pérez, N., Castillo, A., López-López, J. A., Pina, V., Puga, J. L., Campoy, G., ... & Fuentes, L. J. (2018). Computer-based training in math and working memory

improves cognitive skills and academic achievement in primary school children: Behavioral results. *Frontiers in Psychology,* 8, 2327.

Sandeep, S., Shelton, C. R., Pahor, A., Jaeggi, S. M., & Seitz, A. R. (2020). Application of Machine Learning Models for Tracking Participant Skills in Cognitive Training. *Frontiers in Psychology*, 11, 1532. https://doi.org/10.3389/fpsyg.2020.01532

Sasser, T. R., Bierman, K. L., Heinrichs, B., & Nix, R. L. (2017). Preschool intervention can promote sustained growth in the executive-function skills of children exhibiting early deficits. *Psychological Science*, *28*(12), 1719-1730.

Sauce, B., & Matzel, L. D. (2013). The causes of variation in learning and behavior: why individual differences matter. *Frontiers in psychology*, 4, 395.

Saucier, G. (1994). Mini-Markers: A Brief Version of Goldberg's Unipolar Big-Five Markers. *Journal of Personality Assessment*, 63(3), 506–516. https://doi.org/10.1207/s15327752jpa6303_8

Schmitt, S. A., Geldhof, G. J., Purpura, D. J., Duncan, R., & McClelland, M. M. (2017). Examining the relations between executive function, math, and literacy during the transition to kindergarten: A multi-analytic approach. *Journal of Educational Psychology*, *109*(8), 1120.

Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental science*, *20*(3), e12372.

Schölkopf, B., Tsuda, K., & Vert, J. P. (2004). Kernel methods in computational biology. *MIT press*. DOI: https://doi.org/10.7551/mitpress/4057.001.0001

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist,* 50(2), 138-166.

Scott, W. A. (1962). Cognitive complexity and cognitive flexibility. *Sociometry*, 405-414.

Settles, B., & Meeder, B. (2016, August). A trainable spaced repetition model for language learning. In Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers) (pp. 1848-1858).

Shanley, L., Clarke, B., Smolkowski, K., Doabler, C. T., Kurtz-Nelson, E. C., & Fien, H. (2021). Examining the Role of Domain–General Skills in Mathematics Learning and Intervention Response in Kindergarten. *Learning Disabilities Research & Practice*, *36*(4), 330-352.

Shawn Green, C., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., ... & Witt, C. M. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. *Journal of Cognitive Enhancement*, *3*, 2-29.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child development*, 75(2), 428-444.

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental science*, 11(5), 655-661.

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of educational psychology*, 101(3), 545.

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do "brain-training" programs work?. *Psychological science in the public interest*, *17*(3), 103-186.

Sigmundsson, H., Englund, K., & Haga, M. (2017). Associations of physical fitness and motor competence with reading skills in 9-and 12-year-old children: a longitudinal study. *SAGE Open*, 7(2), 2158244017712769.

Smid, C. R., Karbach, J., & Steinbeis, N. (2020). Toward a science of effective cognitive training. *Current Directions in Psychological Science*, 29(6), 531-537.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 13–59). W H Freeman/Times Books/ Henry Holt & Co.

Söderqvist, S., & Bergman Nutley, S. (2015). Working memory training is associated with long term attainments in math and reading. *Frontiers in psychology*, 6, 1711.

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early childhood research quarterly*, 19(1), 99-120.

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, 110(45), 18116-18120.

Steyvers, M., & Schafer, R. J. (2020). Inferring latent learning factors in large-scale cognitive training data. *Nature Human Behaviour*, 4(11), 1145-1155.

Sutin, A. R. (2017). Openness. In T. A. Widiger (Ed.), The Oxford handbook of the Five Factor Model (pp.83–104). *Oxford University Press*.

Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of educational psychology*, 96(3), 471.

Taatgen, N. A. (2001, July). A model of individual differences in learning air traffic control. In *Proceedings of the fourth international conference on cognitive modeling* (pp. 211-216). Mahwah, NJ: Erlbaum.

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., Chang, P., Pollard, K., Lala, N., Alvarez, G. A., & Gabrieli, J. D. E. (2013). Failure of working memory training to enhance cognition or intelligence. *PloS One*, 8(5), e63614. https://doi.org/10.1371/journal.pone.0063614

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Titz, C., & Karbach, J. (2014). Working memory and executive functions: effects of training on academic achievement. *Psychological research*, 78(6), 852-868.

Traut, H. J., Guild, R. M., & Munakata, Y. (2021). Why does cognitive training yield inconsistent benefits? A meta-analysis of individual differences in baseline cognitive abilities and training outcomes. *Frontiers in Psychology*, 12, 662139. DOI: https://doi.org/10.3389/fpsyg.2021.662139

Tullo, D., Feng, Y., Pahor, A., Cote, J. M., Seitz, A. R., & Jaeggi, S. M. (2023). Investigating the Role of Individual Differences in Adherence to Cognitive Training. *Journal of Cognition*, *6*(1).

Vanbinst, K., & De Smedt, B. (2016). Individual differences in children's mathematics achievement: The roles of symbolic numerical magnitude processing and domain-general cognitive functions. *Progress in brain research*, *227*, 105-130.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*, 1413-1432.

Vermeir, J. F., White, M. J., Johnson, D., Crombez, G., & Van Ryckeghem, D. M. (2020). The effects of gamification on computerized cognitive training: systematic review and meta-analysis. *JMIR serious games*, *8*(3), e18644.

Von Bastian, C. C., Langer, N., Jäncke, L., & Oberauer, K. (2013). Effects of working memory training in young and old adults. *Memory & cognition*, 41(4), 611-624.

Von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory training: A review. *Psychological Research*, 78(6), 803–820. DOI: https://doi.org/10.1007/s00426-013-0524-6

Vygotsky, L. S. (1978). Mind in Society: the Development of Higher Psychological Processes. *Cambridge, MA: Harvard University Press*.

Wang, C., Jaeggi, S. M., Yang, L., Zhang, T., He, X., Buschkuehl, M., & Zhang, Q. (2019). Narrowing the achievement gap in low-achieving children by targeted executive function training. *Journal of Applied Developmental Psychology*, 63, 87-95.

Wang, Y., Guo, C., Chu, Y., Hwang, J. N., & Feng, C. (2020). A cross-domain hierarchical recurrent model for personalized session-based recommendations. *Neurocomputing*, 380, 271-284.

Ward, N., Hussey, E. K., Cunningham, E. C., Paul, E. J., McWilliams, T., & Kramer, A. F. (2019). Building the multitasking brain: An integrated perspective on functional brain activation during task-switching and dual-tasking. *Neuropsychologia*, 132, 107149.

Waris, O., Jaeggi, S. M., Seitz, A. R., Lehtonen, M., Soveri, A., Lukasik, K. M., Söderström, U., Hoffing, R. A. C., & Laine, M. (2019). Video gaming and working memory: A large-scale cross-sectional correlative study. *Computers in Human Behavior*, 97, 94–103. https://doi.org/10.1016/j.chb.2019.03.005

Wechsler, D. (1991). Manual for the Wechsler Intelligence Scale for Children-Third Edition. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). Wechsler Memory Scale 3rd ed The Psychological Corporation (86).

Whitehurst, G. J., Arnold, D. S., Epstein, J. N., Angell, A. L., Smith, M., & Fischel, J. E. (1994). A picture book reading intervention in day care and home for children from low-income families. *Developmental psychology*, 30(5), 679.

Wiemers, E. A., Redick, T. S., & Morrison, A. B. (2019). The Influence of Individual Differences in Cognitive Ability on Working Memory Training Gains. *Journal of Cognitive Enhancement*, 3(2), 174–185. https://doi.org/10.1007/s41465-018-0111-2

Williams, B. R., Ponesse, J. S., Schachar, R. J., Logan, G. D., & Tannock, R. (1999). Development of inhibitory control across the life span. *Developmental psychology*, 35(1), 205.

Williams, R., Citkowicz, M., Miller, D. I., Lindsay, J., & Walters, K. (2022). Heterogeneity in Mathematics Intervention Effects: Evidence from a Meta-Analysis of 191

Randomized Experiments. *Journal of Research on Educational Effectiveness*, *15*(3), 584-634.

Willmott, C. J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, 63(11), 1309-1313.

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology*, 5. https://doi.org/10.3389/fpsyg.2014.00213

Xie, Y., Near, C., Xu, H., & Song, X. (2020). Heterogeneous treatment effects on children's cognitive/non-cognitive skills: a reevaluation of an influential early childhood intervention. *Social science research*, *86*, 102389.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zhang, Z. (2016). Multiple imputation with multivariate imputation by chained equation (MICE) package. *Annals of translational medicine*, 4(2).

Zhang, R. Y., Chopin, A., Shibata, K., Lu, Z. L., Jaeggi, S. M., Buschkuehl, M., ... & Bavelier, D. (2021). Action video game play facilitates “learning to learn”. *Communications biology*, 4(1), 1-10.

Zhang, Q., Wang, C., Zhao, Q., Yang, L., Buschkuehl, M., & Jaeggi, S. M. (2019). The malleability of executive function in early childhood: Effects of schooling and targeted training. *Developmental science*, 22(2), e12748.

Zinke, K., Zeintl, M., Eschen, A., Herzog, C., & Kliegel, M. (2012). Potentials and limits of plasticity induced by working memory training in old-old age. *Gerontology*, 58(1), 79-87.

Zinke, K., Zeintl, M., Rose, N. S., Putzmann, J., Pydde, A., & Kliegel, M. (2014). Working memory training and transfer in older adults: Effects of age, baseline performance, and training gains. *Developmental Psychology*, 50(1), 304–315. https://doi.org/10.1037/a0032982