Alternating Finger Tracing For Learning: A Cognitive Load Theory Perspective

Xufei Zhang

A thesis submitted to fulfil the requirements of the degree of Doctor of Philosophy

Faculty of Arts and Social Sciences

The University of Sydney

6th February 2026

Statement of originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged.

Signature:

Name: Xufei Zhang

Date: 6th February 2026

Statement of Gen AI Attribution

No content produced by generative AI tools has been used in the preparation of this thesis.

Signature:

Name: Xufei Zhang

Date: 6th February 2026

Acknowledgements

At this very moment, it is raining. Sydney is not a city always rain. The rain feels like mirror the doctoral journey spanning over three years — now falling, now ceasing, dense yet gentle. I am most grateful that such an experience has graced my life, just as I truly love rainy days like today.

This wonderful journey would not have been possible without my supervisors' constant companionship and unwavering support. My leading supervisor Dr Paul Ginns has inspired me greatly at every stage of experimental design, implementation, and analysis. He steadfastly taught me, through his own example, how to approach research with rigorous and critical thinking. His very existence embodies my idealised vision of a scholar devoted to scientific research, profoundly influencing me over these past three years. My associate supervisor Dr Jennifer Way and Dr Bronwyn Reid O'Connor have always given me tremendous encouragement and provided numerous suggestions in detail. I am truly grateful for having made the decision to pursue my research in Cognitive Load Theory with Paul, and I am really appreciating such an exceptional supervisor team companioning me throughout these past three years.

I extend my gratitude to all students who participated in the experiments; it is through your involvement that this research has been successfully conducted. I am sincerely grateful to Dr. Mei Wang and Dr. Heping Xie, who assisted me with experimental site and other necessary support in conducting this time-consuming experiment in China.

A special thank goes to my partner, Binglun Li, who has been by my side through every moment of immense pressure and emotional breakdown. I also want to thank my parents, Saijun Xu and Zhiqiang Zhang, who are the best parents in the world, allowing their daughter to continue as a full-time student for many carefree years. I would like to dedicate this PhD thesis to my beloved family, that they may share in this grateful achievement with me.

Contents

Abstract

Finger tracing is an effective instructional method to support cognitive learning. Drawing on cognitive load theory, its effect has been demonstrated across various subjects (e.g., mathematics, physiology), learner demographics (children, adults), and instructional designs (worked examples, expository texts with diagrams). The variability of tracing – operationalised as alternating which finger traces across multiple steps in a worked example – has the potential to further support learning. Three experiments were conducted to explore the effects of alternating tracing on mental math learning, intrinsic motivation, and cognitive load.

In Experiment One, 115 university students from Shanghai, China, were randomly assigned to either non-alternating-finger or alternating-finger tracing conditions to learn a mental math strategy. Data from 95 fully compliant participants were included in the analysis. Participants' pre-test (calculation fluency test) performance was included as a covariate, representing prior calculation ability. Contrary to hypotheses, the non-alternating group performed better than the alternating group across practice questions and immediate post-test questions. Repeated-measure analysis of covariance on immediate and delayed post-test scores found a significant interaction between condition and test phase on post-test scores, suggesting the alternating group showed greater improvement across two testing phases. An interaction between prior calculation ability and condition was found on intrinsic cognitive load, suggesting participants whose pre-test scores were lower than 64 perceived lower intrinsic cognitive load in the single condition compared to the alternating condition. Participants in the alternating condition reported higher extraneous cognitive load. No

significant differences were found on intrinsic motivation or germane processing.

In Experiment Two, 112 university students from Guangdong, China, were recruited. Data from 100 fully compliant participants were analysed. One of the 100 participants did not come for the delayed test on the next day, but this data was still kept for analyses not requiring delayed-test data. Procedures were identical to Experiment One, except for adjustments of specific numbers in the learning materials. No differences between conditions were found on any performance measures. Repeated-measure analysis of covariance found better performance in the delayed post-test, but no difference between conditions. The alternating group again reported higher intrinsic and extraneous cognitive load. Intrinsic motivation, and germane processing did not differ between groups.

In Experiment Three, 108 university students from Shanghai, China, were recruited. Ninety-seven participants were fully compliant with the instructions, and one did not come to the delayed test. Procedures were adjusted: learning was divided into two stages, with three worked examples and a self-report questionnaire in each stage. All participants applied the non-alternating finger tracing strategy in the first stage and traced according to the condition in the second stage. Immediate and delayed post-tests included similar questions and transfer questions. Participants in two conditions had similar performance in first-stage learning, indicating they were equivalent in learning ability. The time to solution on first-stage practice questions was included as another covariate, representing participants' prior learning ability. Controlling for prior calculation ability and learning ability, no significant effect was found between conditions on participants' general performance in second-stage practice learning and immediate post-test. However, marginal effects were found on immediate transfer scores and delayed total scores, suggesting the alternating group had better performance. Repeated-measure analysis of covariance found marginal main effects of condition on total scores, similar question's accuracy, and transfer question's accuracy, suggesting the alternating group performed better. Repeated-measure analysis of variance or analysis of covariance was applied to check participants' intrinsic motivation and cognitive load in two stages of learning. An interaction effect was found between the learning stage and the condition on intrinsic motivation: the non-alternating group demonstrated a more substantial decline in intrinsic motivation across two stages, suggesting the potential motivation retention effect of the alternating tracing strategy. The interaction effect was also marginally found on extraneous cognitive load, suggesting non-alternating group had lower extraneous cognitive load the alternating group had higher extraneous cognitive load in the second stage.

Furthermore, a new research question - how does variation in tracing actions affect learning performance? - was addressed and analysed as exploratory research. This question was raised when video coding analyses to check compliance revealed substantial variation in the number of tracing actions actually made by participants. Participants excluded from the above analyses on the basis of non-compliance (insufficient tracing actions) were included in the dosage analyses, as they represent the natural variation in tracing actions. Thus, 110, 109, and 106 participants from Experiments 1, 2 and 3 were included in the dosage analyses. A general positive correlation between tracing dosage and learning performance was found in Experiments 1 and 2. However, this effect was not found in Experiment 3, potentially due to a ceiling effect on practice scores. Tracing's dosage effect on learning performance demonstrated a declining trend over time. No correlation was found between tracing dosage and various cognitive load.

Overall, the results of the current study indicated that alternating fingers tracing may enhance learning when cognitive load is suitably managed, as in Experiment 3. Though higher extraneous cognitive load was experienced in Experiment 3, participants showed better performance, especially on the delayed post-test and transfer questions, suggesting that the alternating fingers tracing strategy may foster deeper information processing. Recommendations for further research on alternating tracing actions are made.

Chapter One: Introduction

1.1 Research Background

Learning constitutes a perennial subject of exploration in human development. Since the very emergence of humankind, diverse modalities of learning have been a close companion to human beings' development. Researchers from neuroscience, cognitive science, behavioural science, psychology, pedagogy and many other fields are all working to uncover the mysteries of cognitive processes. From "mimetic" motor acts to "mythic" spoken language (Donald, 1991), from simple phonetic expression to a completed symbolic system, no matter how the content and forms of human learning change, the pursuit of more effective cognitive learning remains constant.

As Donald (1991) summarised, today's human mind is a hybrid structure that is partly built from vestiges of earlier biological stages as well as new external memory devices that altered its organisation. Humans still initiate their cognition of the world through tactile, auditory, and visual perception, the most fundamental sensory modalities that rely on biological structure. However, the perceived world grows increasingly complex as the individual matures, acquiring language and abstract symbolic systems.

Aligning with this argument of human beings' learning features, evolutionary psychologist David Geary (2008) posited two fundamentally distinct categories of knowledge: *biologically primary knowledge* and *biologically secondary knowledge*. The former category refers to the knowledge or skills that humans have evolved to acquire effortlessly (e.g., learning a gesture, speaking a "mother tongue"). The knowledge belonging to the latter category (e.g., reading, writing, science, and mathematics) requires extended conscious effort to learn. As biologically secondary knowledge has developed only relatively recently in humankind's history as a species (e.g., writing has only been in existence for a few thousand years), humans are not evolved to master biologically secondary knowledge as effortlessly as biologically primary knowledge (Sweller et al., 2011). As biologically secondary knowledge does not directly support or undermine humans' survival and evolution, we have neither the motivational impetus nor the genetically inspired ability to assimilate biologically secondary information automatically. To acquire biologically secondary knowledge, people need to consciously allocate their working memory resources to the acquisition because humans have not evolved to know how this type of knowledge should be processed (Paas & Sweller, 2012; Sweller, 2008).

As a major contemporary theory of instructional design, the current iteration of Cognitive Load Theory (CLT) is substantially informed by the above evolutionary perspectives (cf. Paas & Sweller, 2012; Sweller, 2008; Sweller et al., 2019). It analogises the evolution of humans' cognitive structures to the evolution of genetically based biological structures, arguing that cognitive learning may be limited by working memory capacity, which is largely determined by the biological features of human beings (Sweller, 2004). CLT researchers have historically been concerned with supporting the learning of biologically secondary knowledge through instructional design, but have more recently considered the interplay between these two categories when people learn.

People acquire most biologically primary knowledge at a very young age. As a result, biologically primary knowledge plays a cornerstone role in most aspects of human cognition (Sweller et al., 2011), supporting the acquisition of biologically secondary knowledge.

Research on the human movement effect (e.g., Ayres et al., 2009; Wong et al., 2009) provided empirical evidence that human movement, as a type of biologically primary knowledge, could facilitate biologically secondary knowledge's learning, finding that when learning hands-on tasks, a dynamic representation is more effective than a static representation because the ability to observe and copy human movement is the biologically primary knowledge humans have evolved to gain. Paas and Sweller (2012) argued that humans' innate ability to act in and perceive the world allows for the automaticity of motor and perceptual resonance in cognitive tasks.

Another similar argument regarding biologically primary knowledge supporting biologically secondary knowledge comes from the perspective of embodied cognition. This theory assumes that all cognitive processes are based on action and perception instead of abstract constructs (Barsalou, 1999), supporting cognition through interactions with the environment (Barsalou, 2008; Rueschemeyer et al., 2009). Research in various domains, including action semantics (Lindemann et al., 2006), language comprehension (Zwaan & Taylor, 2006) and neuroscience (Glenberg et al., 2008), has provided biological and psychological evidence for embodied cognition.

Drawing from the theoretical perspectives mentioned above, one recent line of CLT-based research focuses on exploring how *tracing* (touching a surface with a finger in an active movement) can support learning. This activity can be traced back to over a century ago when *Sandpaper Letters* were invented by Maria Montessori. In this activity, students trace letters' cut-outs made of sandpaper with their fingers in the same sequence as writing the letter, with the teacher pronouncing the letter's sound at the same time (Montessori, 1912).

Subsequent research has demonstrated that tracing can support learning in a range of domains across various age groups.

The effect of tracing appears to be robust. However, there is still much to learn about its underlying mechanisms and boundary conditions for effectiveness (Park et al., 2019). Researchers have begun to focus on potential compound effects (Sweller et al., 2019, p. 270) of tracing, referring to "effects that alter the characteristics of other simple cognitive load effects". Conducting compound effects research is a trend in CLT, as such research identifies other cognitive load effects' boundary conditions and thus promotes CLT's development as a theory. Wang et al. (2022) combined another cognitive load effect, the imagination effect (Cooper et al., 2001), with finger tracing and found that sequencing tracing and imagination supported learning to a greater extent than an equivalent period spent on tracing only.

Tracing's potential to support learning is far from being tapped. The current research speculates that the variability effect (Paas & Van Merriënboer, 1994) has similar potential to the imagination effect in being a cognitive load effect that can compound the effect of tracing. In this thesis, combining tracing with the variability effect is operationalised as alternating two index fingers that are used to trace a given step in a worked example. A series of experiments was conducted to determine whether alternating fingers tracing affects students' cognitive load and subsequent learning compared to classic non-alternating finger tracing.

1.2 Significance

Considering the future development potential of finger tracing on learning, as well as the yet-to-be-clarified boundary conditions of how finger tracing supports learning, comparing non-alternating finger tracing (where the tracing effect has been robustly demonstrated) and alternating finger tracing (the novel tracing strategy this study targets at) is a necessary and worthwhile step to take. Therefore, the present study is significant for its role in continuing explore the tracing effect and its implementation in educational practice.

On a theoretical level, the present study expands the scope of finger tracing and cognitive load research. By creating the alternating finger tracing strategy via incorporating the tracing effect and variability effect together, this research explores the effect of alternating finger tracing on learning performance, intrinsic motivation, and various types of cognitive load in detail, contributing to tracing effect, embodied cognition, cognitive load theory, and variability research.

On a practical level, this research investigates a potentially useful learning strategy to support learning. As an easily implemented learning strategy, alternating finger tracing could also offer students, parents, teachers, and educators a simple instructional intervention way to support learning. The present study also suggests a tracing sequence that combining non-alternating tracing and alternating, enabling finger tracing to better accommodate learners of varying abilities.

1.3 Structure of the Thesis

This chapter briefly described the research background of finger tracing effect from evolutionary psychology and CLT perspective, and proposed idea of incorporating variability effect to constitute alternating tracing.

Chapter Two provides a comprehensive description of Cognitive Load Theory, beginning with a general overview on human cognitive architecture, followed by discussions on different types of cognitive load and their relationship to intrinsic motivation. The following section introduces six cognitive load effects relevant to the present study, particularly the variability effect. The last section of this chapter focuses on the incorporation of evolutionary perspective into CLT, stressing the potential of finger tracing as a type of biologically primary knowledge that could support biologically secondary knowledge learning. Chapter Three further discusses the role of finger tracing in promoting learning in detail. This chapter begins by reviewing the empirical studies on finger tracing, demonstrating its applicability across various subject domains, populations, and learning contexts. Then, potential mechanisms of the tracing effect are discussed. Finally, recent research trends on compound tracing effects and tracing dosage are discussed, concluding that tracing research still has considerable scope for development. Thus, the present study constitutes an investigation of a potential compound tracing effect by incorporating the tracing and variability effects.

Chapter Four discusses how variability should be incorporated into finger tracing. The characteristic of variability effect are reviewed, focusing on its interaction with other cognitive load effects. The theoretical possibilities for combining variability effect and finger tracing effect are then discussed. The last section of this chapter discusses options for combining the variability and tracing effects, including the operationalisations investigated in the present study. Chapter Five outlines the methodology employed in this study. The first section establishes that the experimental method constitutes the fundamental approach of this research, and briefly introduced the aim of the present study. Based on the research findings reviewed in previous chapters, the present study proposes hypotheses regarding how alternating finger tracing respectively influences learning performance, intrinsic motivation, and types of cognitive load. The next section discusses certain designs in the experiments in present study, including designs concerning test time to solution and delayed post-test.

Chapter Six to Eight introduce the three experiments in the present study in detail, including methods, results, and discussion for each experiment. Each experiment builds upon the preceding one, with adjustments made to the experimental procedures and materials. These modifications are detailed in the corresponding sections. Experiment 3 involved a significant adjustment to the procedure, splitting the learning phase into two distinct stages. Given the adjustments to variables, the hypotheses have consequently been modified accordingly. Thus, Chapter Eight also includes the discussion and presentation of the adjusted hypotheses. Chapter Nine presents the supplementary research on tracing dosage effect. This supplementary research arose from findings on substantial variations in individual actual tracing actions during the processing of video coding data for formal experiments.

Chapter Ten summarizes major findings of three experiments, discusses the theoretical and methodological implications of the present study, and points out the limitations of the present study, accompanying by the suggestions for future research.

Chapter Two: Cognitive Load Theory

2.1 Theory Overview

Cognitive load theory (CLT) is an educational theory that was initially proposed by John Sweller (Sweller, 1988) and continuously developed by Sweller and other researchers (Chandler & Sweller, 1991; Kalyuga & Plass, 2025; Paas et al., 2003, 2004; Sweller, 1994, 2003, 2022, 2023; Sweller et al., 1998, 2011, 2019). CLT is predicated on human cognitive architecture and its role in instructional design. It focuses on how cognitive load, "the relative demand imposed by a particular task in terms of mental resources required" (American Psychological Association, 2018a), influences learning.

As learning (American Psychological Association, 2018b) and memory (American Psychological Association, 2018c) are two closely related concepts insofar as learning's final outcomes are preserved and presented by memory, understanding the structure of human memory has therefore become a necessary part of learning research. CLT considers two memory stores of human cognitive architecture: a working memory limited in capacity and duration (Cowan, 2001; Peterson & Peterson, 1959), standing in marked contrast to long-term memory with no known upper limit (Baddeley, 1997; Sweller et al., 2011) and a relatively permanent store (Atkinson & Shiffrin, 1968). From an instructional standpoint, a limited working memory capacity means that only a limited cognitive load can be imposed before learning is hindered. However, too low a cognitive load means there are under-utilised working memory resources, which also could make learning inefficient. Teachers should take working memory's limitation into account when instructing, as it is a "bottleneck" for learning (Paas & Van Merriënboer, 2020). Thus, the broad goal of research informed by CLT is to identify instructional design principles that keep the cognitive load at an optimal level.

Beyond understanding the process of human cognitive learning, CLT holds that comprehending the object of learning, and more precisely, alternative categories of knowledge, is equally crucial. Incorporating evolutionary theorising, Geary (2008) posited two fundamentally distinct categories of knowledge: *biologically primary knowledge* and *biologically secondary knowledge*. The former category refers to the knowledge or skills that humans have evolved to acquire effortlessly (e.g., learning to gesture, speaking a “mother tongue”). The knowledge belonging to the latter category (e.g., reading, writing, science, and mathematics) requires extended conscious effort to learn. CLT researchers have historically been concerned with the development of biologically secondary knowledge, but have more recently considered the interplay between these two categories when people learn.

This chapter presents CLT’s core ideas, concepts and findings in detail, starting from human cognitive structure and working memory limits. Then, three types of cognitive loads and related research are introduced. Following that, this chapter lists several classic cognitive load effects and then moves to CLT’s new trends over the last decade, especially from an evolutionary perspective. Two categories of knowledge are introduced, followed by empirical research discussing how biologically primary knowledge (e.g., hand movements, gestures, pointing and tracing) supports biologically secondary knowledge learning.

2.2 Human Cognitive Architecture and Working Memory Limit

As Sweller (2012) articulates, cognitive load theory is a theory that uses knowledge of human cognitive architecture to generate instructional procedures. During CLT’s development and formation, it integrated various theories and models on learning, memorising, and information processing, including short-term memory discussing temporary information storage (Miller, 1956), working memory discussing information storage and processing (Baddeley & Hitch, 1974), the multi-store memory model discussing memory structure (Atkinson & Shiffrin, 1968), and long-term working memory discussing expert performance (Ericsson & Kintsch, 1995). Based on basic research on human cognitive architecture, CLT generates its core argument that learning effectiveness is limited by working memory capacity, especially when processing new information. These limitations can be markedly reduced when a schema (a cognitive construct that permits to treat multiple elements of information as a single element) is formed in long-term memory so that the previously learned material can act as a central executive, and working memory becomes maximally effective (Sweller, 2003). This section reviews research on working memory and long-term memory, and subsequent scholarship on better utilising memory processing mechanisms to support learning.

2.2.1 Working Memory

Research on working memory can be traced back to Miller's (1956) findings on the "magical number seven, plus or minus two". Miller argued that there was a clear and definite limit of seven items on human short-term memory, which he called "span of immediate memory". Although contemporary researchers (e.g., Aben et al., 2012; Colom et al., 2006) tend to regard short-term memory and working memory as two similar yet distinct concepts with different emphases, the proposal of the "7±2" memory capacity limit still laid the foundation for subsequent research on working memory.

The concept of "working memory" was first introduced in Miller et al.'s (1960) book, where they described working memory as the memory that humans use for the execution of plans as a kind of quick access. This statement provided the earliest definition of working memory from the functional perspective.

Atkinson and Shiffrin (1968) then used the concept in their multi-store memory model. In the multi-store memory model, memory was divided into three structural components: the sensory register, the short-term store (STS), and the long-term store (LTS). This categorisation also set the stage for later researchers to discuss short-term memory (or temporary memory, immediate memory) versus long-term memory. Atkinson and Shiffrin regarded the STS as the subject's working memory, receiving selected inputs from both the sensory register and the long-term store. In Atkinson and Shiffrin's model, the STS can be understood as a container, and the working memory is its contents. Atkinson and Shiffrin argued that the information in the short-term store may exist for about 30 seconds without rehearsal (cf. Peterson & Peterson, 1959), indirectly defining the duration of working memory existence.

Atkinson and Shiffrin (1971) further pointed out that the STS took the role of a controlling executive system in which decisions were made, problems were solved, and information flow was directed. However, Baddeley and Hitch (1974) found there was little empirical evidence supporting STS's control function and thus questioned Atkinson and Shiffrin's model on whether STS actually played this role in information processing. To figure out whether reasoning, comprehension, and learning all share a common working memory system, and what the relationship is between this system and short-term store, Baddeley and Hitch (1974) conducting multiple experiments with different combinations of three types of learning task (verbal reasoning, language comprehension, and free recall of unrelated words) and three types of interference task (memory load, phonemic similarity, and articulatory suppression). These dual-task experiments found that participants' performance on each of two tasks (one learning task and one interference task) under the dual-task condition can reach a similar level of performance under the single-task condition when two tasks come from different domains (e.g., verbal tasks and image tasks). After comparing and summarising all these experiment results, Baddeley and Hitch (1974) proposed a working memory system that consisted of a limited capacity "work space" that can be divided between storage and control processing demands. The working memory system can be viewed as a modification of Atkinson and Shiffrin's multi-store memory model, but focusing more on the information processing tasks rather than the memory system. In Baddeley and Hitch's (1974) model, the working memory represented a control system with limits on both storage and processing capabilities.

Baddeley (1986) further refined the working memory model and described it in detail, which has come to be known as the multicomponent working memory model. Three specialised components (central executive, phonological loop, and visuo-spatial sketchpad) constitute this multicomponent working memory system. The first component is the central executive; it is a supervisory system that takes the responsibility to control and regulate the working memory system. Specifically, it undertakes functions like coordinating the other two systems, allocating and switching attention, and activating representations within long-term memory (Baddeley & Logie, 1999). The central executive is not involved in temporary storage, which is the function of the other two components. The phonological loop and visuo-spatial sketchpad are specialised temporary memory systems; the former one refers to the phonologically based store and the latter one refers to the visuospatial store.

The multicomponent working memory model (Baddeley, 1986) answered the question of why learning tasks from two different domains did not interfere with each other, consistent with the experimental results (Baddeley & Hitch, 1974). However, in the multicomponent working memory model, verbal and nonverbal information is stored in two isolated systems, which leaves the question of how the two types of information integrate.

To answer this question, Baddeley (2000) further revised the model and introduced another component, the episodic buffer, into the model. The episodic buffer is another storage system supervised by the central executive; it provides temporary storage and is capable of integrating and storing information from various sources (Baddeley, 2000). The episodic buffer is argued to provide a temporary interface between the phonological loop, the visuo-spatial sketchpad and the long-term memory.

The four-component working memory model clearly explains how working memory works on information processing and its relationship to long-term memory, consistent with the empirical results at that time. However, as more and more research from the embodied cognition perspective found that body movements can support learning (e.g., Glenberg et al., 2008; Lindemann et al., 2006; Zwaan & Taylor, 2006), Baddeley's (2000) three specific information storage components can no longer explain the phenomena. Baddeley (2012) responded to those questions and further adjusted the multicomponent working memory model, adding another information channel, the haptic channel, into the visuo-spatial sketchpad. The haptic channel is responsible for the tactile and kinaesthetic coding that comes from touching objects and moving body parts.

Baddeley's multicomponent working memory model was not the only model proposed during those years. Cowan's embedded-processes model (Cowan, 1988, 1995, 1999) holds the opposite opinion as the multicomponent working memory model, arguing that rehearsal and retrieval processes actually work upon activated traces of long-term memory, not the separated representations held in the temporary storage in working memory. Memory is held to be a unitary construct but with three distinct levels of activation, including: a) inactive representations of long-term memory, b) a subset of activated representations of long-term memory that is considered as the short-term store; c) a subset of activated representations in the short-term store that is in the focus of attention and awareness (Spillers et al., 2012). In the embedded-processes model, working memory is not viewed as an isolated store, but an embedded field where three levels of memory are hierarchically processed (Cowan, 1999). Compared to the multicomponent working memory model, the embedded-processes model better explained how prior knowledge stored in long-term memory is used for new information processing, with the assumption that working memory and long-term memory are structurally intertwined (Schweppe & Rummer, 2014).

In addition to an alternative view of WM structure, Cowan (2001, 2005) also proposed an alternative perspective on working memory capacity. From the embedded-processes model, the limit of working memory capacity is essentially the limit of attentional resources (Cowan, 1988). Cowan (2001) argued that the "magical number" of short-term memory is four, not seven. Cowan (2001) set very strict conditions to observe the single, central capacity limit and found that the mean memory capacity in adults is three to five chunks. Cowan (2001) referred to “7±2” as the compound short-term memory limit that was suitable for materials where the exact number of chunks was hard to ascertain, and proposed “4 chunks” as the pure short-term memory limit that was more widely used.

Despite the debates on working memory’s detailed structure and capacity limit, there is a consensus view among researchers that: a) working memory plays a crucial role in information processing; b) working memory interacts with long-term memory; c) working memory has a limited capacity (see Oberauer et al., 2018). These consensus characteristics of working memory underpin the basic idea of Cognitive Load Theory of maximising learning by activating and controlling cognitive load at an optimal level.

2.2.2 Long-Term Memory

In contrast to the limited capacity of working memory, there is no clear upper limit to the storage capacity of long-term memory. As Atkinson and Shiffrin (1968) described, human learning starts with receiving information, followed by complicated information processing in working memory, and finally storing it in long-term memory.

Long-term memory’s importance in cognitive tasks was first proposed by de Groot (1965) in research on chess expertise. Chase and Simon (1973) replicated and expanded this line of research, finding that experts performed better on a memory task based on real-game positions than beginners, but not when chess pieces in random positions were to be remembered. This finding suggested that experts’ better performance mostly came from the large number of chunks or patterns stored in their long-term memory, rather than their superb ability or larger working memory capacity. Following research demonstrated similar results in various domains (e.g., Egan & Schwartz, 1979; Gobet & Simon, 1996; Sweller & Cooper, 1985), indicating that long-term memory can generally help cognitive learning.

Considering how prior knowledge stored in long-term memory supports the learning process, the schema construct was proposed by Piaget (1952) to explain how knowledge is stored in long-term memory and how people actively construct their understanding of the world. A schema is defined as "a cohesive, repeatable action sequence possessing component actions that are tightly interconnected and governed by a core meaning" (Piaget, 1952, p. 7). In the context of human cognition and pedagogical instruction, Sweller defined a schema as "a cognitive construct that permits to treat multiple elements of information as a single element, categorised according to the manner in which it will be used" (Sweller, 1999, p. 10). According to schema theory (e.g., Ghosh & Gilboa, 2014; Marshall, 1995), schemas in long-term memory can be quickly activated when individuals encounter a related situation. The schema can be expanded or adjusted according to an individual's experience. As additional information is encoded in specific schemas and complicated connections are established between schemas, the individual becomes an expert in that field. Compared to a novice, experts can more quickly and correctly make connections between the current learning task and existing relevant schemas in long-term memory, resulting in better performance in the learning tasks (e.g., Chi et al., 1981).

Rather than being a simple storage system for knowledge, long-term memory actually played an important role in the learning process (Sweller et al., 1998). Schemas stored in the long-term memory not only can be processed consciously as discussed above, but can also be automatically processed (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) with minimal conscious effort. Sufficient practice can promote schema automation, resulting in a lower working memory load required to complete a given task. For example, adult native speakers can read a word without consciously processing the individual letters, but second language beginners usually need to combine the letters in their minds first to think of the word's meaning and pronunciation. Schema automation is effective in supporting problem solving (Kotovsky et al., 1985) because more working memory capacity is available to be used to solve sophisticated problems when a learner possesses a more automated schema (Sweller et al., 1998).

The discovery of the important role of long-term memory in supporting learning constitutes another fundamental tenet of Cognitive Load Theory: schema construction and automation will support subsequent learning by freeing working memory capacity and avoiding overload (Sweller et al., 1998).

2.3 Cognitive Load and Instructional Intervention Guidelines

Since working memory capacity is limited and indispensable to learning, it is necessary to control the burden on working memory capacity to avoid overload. The concept of cognitive load was thus proposed as a construct representing the load imposed on the cognitive system to perform a learning task (Sweller et al., 1998). The current iteration of CLT (Paas & Van Merriënboer, 2020; Sweller et al., 2019) stipulates two distinct categories of cognitive load: *intrinsic cognitive load* and *extraneous cognitive load*. Intrinsic cognitive load is imposed by the inherent complexity of the information the learner needs to understand, given their prior knowledge, irrespective of the instructional procedures used. In contrast, extraneous cognitive load is imposed not by the information itself but rather by how it is presented or by the activities learners engage in (Paas et al., 2003; Sweller et al., 2011;

Sweller et al., 1998; van Merriënboer & Sweller, 2005). Both intrinsic cognitive load and extraneous cognitive load can be adjusted by instructional interventions depending on their own characteristics.

2.3.1 Intrinsic Cognitive Load

Intrinsic cognitive load directly relates to the inherent complexity of the learning content (Sweller, 1994, 2010; Sweller & Chandler, 1994). It can be viewed as a type of "necessary effort" for a learner to acquire the knowledge. For a specified learning task and specified learner knowledge level, intrinsic cognitive load is fixed, other than changing the learning task or changing the learner's knowledge level. Intrinsic cognitive load can only be altered by changing the nature of what is learned or by the learning activity itself (Sweller, 2010). For a particular task and learner knowledge level, *element interactivity* is assumed to be the key determinant of intrinsic cognitive load. An element is anything that has been or needs to be learned (e.g., concept, procedure) (Sweller, 1994, 2010), while the level of element interactivity reflects the number of interacting elements that must be processed simultaneously in working memory (Sweller, 1994).

Low element interactivity learning tasks allow learners to learn individual elements in isolation with minimal reference to other elements, thus resulting in a low load on working memory. A learning task can be difficult but low in element interactivity, because the task requires many elements to be learned serially rather than simultaneously, without holding too many elements in the working memory at the same time (Sweller, 1994, 2010). For example, learning proper nouns in foreign languages is a low-element interactivity task; a second language speaker who starts to live abroad can learn the word "dermatologist" in the clinic without knowing what the word for “cardiologist” is. The second language speaker may learn the proper nouns of different doctors separately in the future when encountering different illnesses.

High element interactivity learning tasks, on the other hand, consist of elements that highly interact with each other and cannot be learned separately. To understand high element interactivity learning tasks, learners have to hold and process the elements in the working memory simultaneously. Compared to learning nouns, learning grammatical properties in a foreign language is a high element interactivity task because learners must relate each word to the others to successfully learn the task. As another example, mathematics tasks also tend to have a high interactivity level. To learn an equation, learners need to take all the digits and the operational rules into account and process them in working memory at the same time (Sweller, 1994, 2010).

However, the element interactivity level depends on the nature of learning content and the learner’s knowledge level simultaneously (Sweller, 2010; Sweller & Chandler, 1994; Tindall-Ford et al., 1997). As learners continue to learn similar tasks, a schema can be constructed. Once a schema has been constructed, it can act as a single element in working memory, and the interacting elements incorporated within the schema do not need to be considered. Thus, a specific low element interactivity learning task to an expert can be a high element interactivity learning task for a novice.

Since intrinsic cognitive load through element interactivity is determined by an interaction between the nature of the material being learned and the expertise of the learner, knowing the knowledge level of the target population of learners is essential when designing instructional materials (Chen et al., 2023; Sweller et al., 1998; Sweller & Chandler, 1994). Some classic instructional interventions considering the element interactivity level are introduced in the next section.

2.3.2 Extraneous Cognitive Load

Other than the load imposed by the intrinsic complexity of what needs to be learned, working memory also needs to handle the load imposed by instructional procedures that are less than optimal, referred to in CLT as *extraneous cognitive load*. This type of cognitive load is considered "unnecessary effort" because it does not relate to the intrinsic characteristics of the learning content, but is only related to the format that the learning content is presented (Paas et al., 2003; Sweller et al., 2011; Sweller et al., 1998; van Merriënboer & Sweller, 2005). Due to working memory's limited capacity and extraneous cognitive load's learning-unrelated characteristic, historically, the primary concern of CLT has been to design appropriate instructional material to reduce extraneous cognitive load (Sweller, 2003, 2004, 2010).

Though extraneous cognitive load is generally believed to be unnecessary and needs to be reduced, it does not necessarily hinder learning. Whether extraneous cognitive load causes a problem partially depends on the intrinsic cognitive load (Sweller et al., 1998; Sweller & Chandler, 1994). Under CLT's additivity hypothesis (Paas et al., 2003), high extraneous cognitive load plus high intrinsic cognitive load may be critical because the total load may exceed the working memory capacity. In contrast, high extraneous cognitive load with low intrinsic cognitive load, where the total load does not exceed the limit, is thus less harmful. To a certain extent, intrinsic cognitive load and extraneous cognitive load have similar attributes: both of them are true loads that need to be processed in the working memory.

Similarity of intrinsic and extraneous cognitive load is also reflected in the fact that element interactivity is the major source of working memory load underlying both extraneous cognitive load and intrinsic cognitive load (Sweller, 2010). If a learning task's element interactivity can be reduced without changing the nature of what is learned, then the load is extraneous cognitive load. For example, presenting a picture of an animal with its scientific name right in the picture has a lower element interactivity level than presenting its scientific name on the other side of the page. The knowledge a student has to learn is the same in both types of picture presentations, but the element interactivity in the two learning tasks is different. Thus, the load is extraneous in such situations. Similarly, if a learning task's element interactivity cannot be altered except by changing the nature of what is to be learned, then the load is intrinsic cognitive load (Beckmann, 2010). However, whether the load is intrinsic or extraneous depends on what needs to be learned. Even for the same learning material, different learning goals may lead to different judgments of intrinsic and extraneous cognitive load (Sweller et al., 1998). Continuing with the previous example, if knowing that an animal and its scientific name is the goal of learning, then presenting the picture and the name far away may cause extraneous cognitive load. But if the learning goal is to observe the details of the animal picture and draw it, then presenting the picture and the name far away does not cause additional extraneous cognitive load.

In summary, extraneous cognitive load is viewed as a kind of unnecessary cognitive load imposed by nonoptimal instructional procedures; it may hinder learning when the combination of intrinsic and extraneous load is greater than a student's working memory capacity. Though extraneous cognitive load is distinct from intrinsic cognitive load by definition and effects, they share a similar underlying mechanism in element interactivity as the major source of both intrinsic cognitive load and extraneous cognitive load.

2.3.3 Germane Processing

Besides intrinsic cognitive load and extraneous cognitive load introduced above, previous iterations of CLT (e.g., Sweller et al., 1998) also considered 'germane cognitive load', defined as effort contributing to schema construction. However, following criticisms of the germane load construct (Kalyuga, 2011; Sweller, 2010), this construct has been reframed as *germane processing*, "working memory resources that are devoted to information that is relevant or germane to learning" (Sweller et al., 2011, p. 57).

In contrast with the previous two categories of cognitive load, germane processing does not contribute to the total cognitive load. Instead, it contributes more to "…redistributing working memory resources from extraneous activities to activities directly relevant to learning by dealing with information intrinsic to the learning task" (Sweller et al., 2019, p. 264). Unlike intrinsic cognitive load and extraneous cognitive load that must be dealt with by working memory with resources allocated to them (Sweller et al., 2011), germane processing does not constitute an independent source of cognitive load. It purely refers to the working memory resources devoted to dealing with the element interactivity associated with intrinsic cognitive load (Sweller, 2010).

Since the germane processing represents how much working memory resources were devoted to processing intrinsic cognitive load related information, it is closely related to intrinsic cognitive load and extraneous cognitive load. If intrinsic cognitive load is high, germane processing tends to be high because a greater proportion of working memory resources needs to be devoted to dealing with essential learning. If extraneous cognitive load is high, germane processing is more likely to be low because more working memory resources are used to deal with extraneous elements imposed by the instructional procedure rather than the lesson elements that need to be understood and learned (Sweller, 2010).

Considering the different characteristics of the various sources of cognitive load, CLT researchers argue that intrinsic cognitive load needs to be optimised, and extraneous cognitive load needs to be reduced to best facilitate learning (Sweller et al., 2011). With such ideal control of intrinsic cognitive load and extraneous cognitive load, the maximal level of working memory resources can be devoted to germane processing, resulting in maximised learning.

2.3.4 Intrinsic Motivation and Cognitive Load

After years of development and refinement, CLT researchers built a convincing framework for instructional design by clarifying the relationship between element interactivity and various sources of cognitive load. However, the above arguments regarding cognitive load and instructional guidelines have a prerequisite that all working memory resources are devoted to learning. To fulfil this prerequisite, learners' intrinsic motivation is important.

As Sweller (2010) has noted, the element interactivity formulation and the instructional guidelines flowing from it assume student motivation is high, and all available working memory resources are devoted to dealing with intrinsic and extraneous cognitive load. Only with a high and stable motivation will working memory resources be guaranteed to be devoted to dealing with cognitive load during the whole learning process. If a learner's intrinsic motivation reduces, then they may allocate some working memory resources to deal with elements that are unrelated to the learning context (e.g., the sound of birds singing outside the window), and the overall working memory resources available for cognitive learning will decrease. Thus, the instructional guideline to improve learning by reducing extraneous cognitive load, thus freeing up working memory resources for learners to deal with intrinsic cognitive load-related information, will not be effective if learners are unmotivated and do not want to use the freed resources to learn (van Merriënboer & Sweller, 2005).

Though CLT researchers have focused more on cognitive structures and processes during learning, there is longstanding recognition that learning is not only affected by cognitive aspects, but also by motivational and emotional factors (Brünken et al., 2010; Feldon et al., 2019; Schnotz et al., 2009). Based on the fundamental view that increasing motivation can increase the working memory resources devoted to learning tasks and further increase germane processing (Paas et al., 2003), maintaining relatively high motivation during the learning process was considered important. This requirement of high motivation, to some extent, changed the idea that learning must be simplified as much as possible to reduce unnecessary cognitive load. As Schnotz et al. (2009) argued, "learning environments, which cut requirements to the basic, may reduce cognitive load, but might not be optimal with respect to the learner's motivation, because a sub-challenge would be experienced as boring by the learner. " For example, removing unnecessary decorative elements (seductive details; for a meta-analysis, see Sundararajan & Adesope, 2020) in lesson materials is a way to avoid distraction and can reduce extraneous cognitive load, but it may also decrease the level of interest and enjoyment learners feel. Furthermore, learners' motivation may be reduced, resulting in lower persistence in learning and poorer learning outcomes.

Though the inclusion of motivational factors makes instructional design more complex, the general guideline that learners should be supported to devote as many working memory resources as possible to dealing with intrinsic cognitive load related information to promote learning remains unchanged. To summarise, considering both cognitive load and motivational aspects simultaneously is essential in instructional design. CLT researchers aim to balance the complex relationship between motivation and cognitive load, generating and testing appropriate instructional methods that maximise germane processing, and thus promote learning to the greatest extent.

2.4 Key Cognitive Load Effects

Aligning with the instructional guidelines introduced in the previous section, many cognitive load effects were generated when an instructional procedure demonstrated its superiority in facilitating learning compared to a more traditional procedure (Sweller et al., 2011). Some of the effects are attributed to the optimisation of intrinsic cognitive load (e.g., the variability effect), and some are attributed to the optimisation of extraneous cognitive load (e.g., the split-attention effect), both of which result in promoting germane processing. There are also many cognitive load effects that are influenced by various principles, and their effects can be altered (disappear or even reverse) by other so-called compound effects (e.g., compound element interactivity effect). Compound effects usually indicate the limits of other simple cognitive load effects, which is regarded as a further development of CLT underpinning substantial research in recent years (Sweller et al., 2019). This section introduces six cognitive load effects germane to the current research. The variability effect, which is the most relevant effect to this PhD study, is analysed in detail.

2.4.1 The Worked Example Effect

Research into worked examples dates back to the 1950s, when researchers applied the learning-by-example method to investigate the learning of concept formation (e.g., Bruner et al., 1956). The worked example effect, referring to providing a step-by-step solution to a problem and better promoting learners' learning (Sweller et al., 2011), has been researched by CLT since the 1980s (Cooper & Sweller, 1987; Sweller & Cooper, 1985). From the CLT perspective, a worked example provides a clear model for learners to construct an expert-like problem-solving schema in long-term memory, while imposing a relatively low working memory load compared to traditional problem-based learning (Sweller et al., 2011). Traditional problem-solving via means-ends analysis requires the learner to process all relevant elements simultaneously in working memory, placing a heavy burden on working memory. By providing worked examples, all the necessary intrinsic interacting elements are encapsulated together, avoiding the overload of working memory (Carroll, 1994).

Empirical evidence for the worked example effect was initially found in the mathematics and science domains. For example, Sweller and Cooper (1985) researched secondary school students' learning on solving algebra manipulation problems, finding that students spent less time to learn worked examples compared to conventional problems, and had better performance on subsequent questions. Zhu and Simon (1987) further demonstrated that worked examples could facilitate students' mathematical learning in the long term, enabling most students to complete a three-year curriculum via the worked-example method in two years of study. Trafton and Reiser (1993) tested the worked example effect in the adult age group, finding that college students benefited more from worked examples when learning computer programming.

Subsequently, the worked example effect has been widely demonstrated in various domains. In STEM (science, technology, engineering and mathematics) disciplines where well-structured problems predominate, the worked example effect has been extensively researched, including statistics problems (Paas, 1992), geometry problems (Paas & Van Merriënboer, 1994), spatial rotation (Pillay, 1994), and many other specific areas; for a meta-analysis on the use of worked examples in mathematics education, see Barbieri et al. (2023). More recent studies have demonstrated the worked example effect in a range of ill-defined domains, including designer recognition (Rourke & Sweller, 2009), drama understanding (Oksa et al., 2010), legal reasoning (Nievelstein et al., 2013) and second language learning (Kyun et al., 2013).

The worked example effect has been shown to be very robust after years of cognitive load theory-based research. Though the worked example effect originated from the cognitive load theory, the theory itself has been influenced by the research findings associated with the worked examples. Failures to generate worked example effect prompted research on a range of other cognitive load effects (Sweller et al., 2011), which are discussed in the subsequent subsections. The split-attention effect is discussed in the next subsection, which is critical to the effectiveness of worked examples.

2.4.2 The Split-Attention Effect

The split-attention effect arose when some worked examples were found to be ineffective (Tarmizi & Sweller, 1988). The worked examples facilitate learning by decreasing extraneous cognitive load, but in some cases, the worked example itself may generate extraneous cognitive load and become ineffective. This may occur when a worked example consists of multiple sources of information (e.g., text and pictures) that are separately presented in the learning material (Chandler & Sweller, 1991; Sweller et al., 1990; Sweller & Chandler, 1994). From a cognitive load perspective, when learners see one part of the information, they are forced to search for the referents and mentally integrate them to understand the overall content. Both searching and integrating require mental efforts, imposing extra extraneous cognitive load on working memory (Chandler & Sweller, 1991; Sweller et al., 2011; Tarmizi & Sweller, 1988; Ward & Sweller, 1990).

Tarmizi and Sweller (1988) first found the split-attention effect when they found that only worked examples that integrated text and related diagrammatic elements triggered the worked-example effect. Subsequent research also found that presenting all materials in a manual better supports learning compared to presenting information across the manual and computer screen (Chandler & Sweller, 1996; Sweller & Chandler, 1994). Thus, CLT suggested that educators integrate all essential information together when designing the learning materials, in order to avoid unnecessary extraneous cognitive load. Specifically, educators can embed the written text within the diagram and place the explanatory text near the concept to reduce the cognitive load caused by searching and matching (Owens & Sweller, 2008; Sweller et al., 2011; Ward & Sweller, 1990). Besides the spatial split-attention effect mentioned above, Mayer (2001) proposed the temporal contiguity principle, pointing out that presenting all necessary information simultaneously is also important, due to the limited duration of working memory (cf. Peterson & Peterson, 1959).

Ginns (2006) reviewed 50 independent studies and demonstrated the robustness of the benefits from reducing split attention across time or space in various domains. This meta-analysis also found that the effectiveness of the split-attention was moderated by the complexity of learning materials, where learners benefited more from avoiding split-attention in complex learning materials (Ginns, 2006). This finding is consistent with the CLT and many other cognitive load effects that the effect tend to be more obvious when intrinsic cognitive load is high (Sweller et al., 2011).

Though numerous studies have demonstrated that integrating related learning elements either over time or space is important for learning (Ginns, 2006), there are indeed some situations where it is impossible to present all information together. In this case, alternative methods to overcome the split attention are needed. Kalyuga et al. (1999) found that directing learners' attention by a colour coding system connecting to relevant information could reduce cognitive load. Similarly, directing learners' attention via an explicit link was found to have a similar learning supportive effect to the fully integrated learning materials (Florax & Ploetzner, 2010).

The split-attention effect is one of the most important cognitive load effects, providing the basic instructional design principle that information needs to be presented in a way that eliminates spatial or temporal separation. Given the fact that information needs to be considered simultaneously to learn, reducing the extraneous cognitive load imposed by holding information in the working memory and searching for referents helps learning (Sweller et al., 2011). Because of its direct and clear mechanism flows from cognitive load theory, the split-attention effect has wider implications for other cognitive load effects. The modality effect discussed in the next subsection relies on the split-attention effect.

2.4.3 The Modality Effect

The modality effect is closely related to the split-attention effect, originating from an alternative way to deal with the split-attention effect (Sweller et al., 2011). Rather than struggling to present all relevant elements visually together, engaging both auditory and visual working memory channels is suggested by the modality effect.

Mousavi et al. (1995) first demonstrated the modality effect in geometry learning. They found that presenting geometry diagrams in visual form and geometry statements in auditory form improved learning compared to visual-only presentations. This finding suggests that working memory has partially independent processors for handling visual and auditory material. Tindall-Ford et al. (1997) replicated the modality effect in electrical engineering and interpreted the modality effect from a CLT perspective. Subsequently, the modality effect has been demonstrated in various experiments (e.g., Jeung et al., 1997; Kalyuga et al., 1999; Mayer & Moreno, 1998; Moreno & Mayer, 1999; for a meta-analysis, see Ginns, 2005).

Besides supporting learning via reducing the extra cognitive load caused by split attention, the modality effect also promotes learning by increasing effective working memory capacity (Sweller et al., 1998). Compared with processing all information in visual working memory alone, adding the auditory channel as extra working memory resources allows learners to process more complicated learning information. This also explains why the modality effect is more obvious in high element interactivity materials (Tindall-Ford et al., 1997). High element interactivity materials may overload the visual channel, which can be reduced by using a dual-modality format (Sweller et al., 2011).

The modality effect can also be explained from an evolutionary perspective. As humans evolved the ability to use vision and hearing to identify an object simultaneously, the involvement of the dual sensory system may rely more on biological rather than cognitive factors for coordination, which results in less cognitively demanding and lower cognitive loads when processing information (Sweller et al., 2019). This evolutionary perspective, which holds that humans have biologically evolved to utilise multiple sense to learn, was then extended to the tactile and kinaesthetic sensory channels. The specific details of this argument are discussed in the following sections.

Though the modality effect is found to be stable and robust in various CLT research, simply using audio-visual instructions does not necessarily promote learning. As working memory for novel information is limited and spoken information is transitory, a lengthy or complex spoken text may exceed working memory limits. Such high element interactivity auditory material may overwhelm working memory and override any possible benefits provided by a dual-modality presentation. Findings from Leahy and Sweller (2016) supported this argument. They conducted two experiments to compare visual-only presentation and audio-visual presentation. The first experiment used lengthy text and found the visual-only presentation was superior, while the second experiment divided the same material into smaller segments and found audio-visual presentation group performed better. Thus, careful design in audio-visual instruction to reduce extraneous cognitive load is necessary to obtain the modality effect (Sweller et al., 2011).

In summary, dual-modality presentation can promote learning because it can reduce the extraneous cognitive load caused by visual split attention and bring extra working memory resources by adding another sensory channel (Sweller et al., 2011). As it arises from the split-attention effect, the modality effect has a similar prerequisite that the visual and auditory information must refer to each other and be unintelligible unless they are processed together. Otherwise, the redundancy effect may occur, which is discussed in the next subsection.

2.4.4 The Redundancy Effect

The redundancy effect is closely related to the split-attention effect and the modality effect. Split attention occurs when learners must integrate information from multiple sources to understand it. In contrast, the redundancy effect may occur when multiple sources of information can be understood separately without mental integration, where extra information is unnecessary or duplicated (Sweller et al., 2011).

The redundancy effect was first reported by Miller (1937), finding that children learning how to read nouns better when the words are presented alone rather than in conjunction with pictures. In this case, the picture was the unnecessary information. When the unnecessary information was presented with the essential information together, the elements associated with the unnecessary information are likely to be processed, resulting in an extraneous working memory load (Sweller et al., 2011). Besides presenting unnecessary information to learning being considered redundant, another common form of redundancy occurs when the same information is presented from multiple resources. This kind of redundancy is caused by duplicated information, including presenting the same information in different forms (for example, the diagram and the text providing the same information), and presenting the same information in different modalities (for example, the identical text being presented via both written and spoken).

Within the cognitive load framework, Chandler and Sweller (1991) first demonstrated the redundancy effect. They conducted several experiments with electrical engineering materials, finding that learners presented with diagram-only instruction spent less time to learn and performed better than those who were presented with the integrated format that combined diagram and text telling the same information. Bobis et al. (1993) replicated the redundancy effect with a paper-folding learning task and found that text-only instruction better promotes learning than text and diagram instruction. To further understand how redundancy affects learners' cognitive load, Chandler and Sweller (1996) used the secondary task method to measure cognitive load and found that redundancy did increase cognitive load.

To summarise, whether it is unnecessary or duplicated information, any additional information not required for schema construction is classified as redundant from the CLT perspective. The redundancy effect occurs when learners involuntarily allocate the working memory resources to process redundant information, which results in increasing extraneous cognitive load and worse learning outcomes (Sweller et al., 2011). Like many other cognitive load effects, the redundancy effect is usually found in learning high element interactivity materials (Sweller & Chandler, 1994), because extraneous cognitive load may not cause harmful results if intrinsic cognitive load is low (Sweller, 1994; Sweller et al., 1998). Element interactivity, which is a key moderator of so many cognitive load effects, is discussed in the next subsection.

2.4.5 The Element Interactivity Effect

As mentioned in previous subsections on sources of cognitive load, the element interactivity is the key source of working memory load underlying both extraneous cognitive load and intrinsic cognitive load. The general finding that CLT-based effects tend to be observed in high element interactivity learning tasks is referred to as the element interactivity effect (Chen et al., 2023; Sweller, 2010; Sweller & Chandler, 1994). The element interactivity effect depends on both intrinsic and extraneous cognitive load (Sweller et al., 2011), and it occurs when a cognitive load effect obtained using high element interactivity information disappears or reverses when using low element interactivity material (Sweller et al., 2019).

In contrast to previous classic cognitive load effects, the element interactivity effect is defined as a compound cognitive load effect, referring to an effect that could alter the characteristics of other cognitive load effects (Sweller et al., 2019). It is treated as a separate cognitive load effect because substantial empirical evidence suggests that the levels of element interactivity have profound effects on other cognitive load effects associated with extraneous cognitive load (Sweller et al., 2011).

The relationship between levels of element interactivity and the effectiveness of various cognitive load effects was found as early as the 1990s (Sweller, 1994), but at that time, it was not defined as a distinct cognitive load effect (Sweller et al., 2019). At that time, the concept of element interactivity was used to explain why some materials were difficult to understand and why extraneous cognitive load only interfered with learning under high cognitive load conditions (Sweller, 1994). As more empirical results emerged, the element interactivity effect had gradually been established as a robust effect with a clear mechanism that directly derives from CLT (Sweller et al., 2011).

As discussed in previous sections, the split-attention effect, the redundancy effect, and the modality effect only tend to be observed when element interactivity is high. Under the additivity hypothesis (Paas et al., 2003), only when element interactivity (and subsequently intrinsic load) is high will those cognitive load effects that work by reducing extraneous cognitive load be obtained. The element interactivity effect has also been implicated in several other compound cognitive load effects (e.g., the expertise reversal effect, the imagination effect) (Sweller et al., 2011).

In summary, the element interactivity effect suggests that most cognitive load effects only manifest when intrinsic cognitive load is high. It is not surprising, as the CLT initially aims to deal with excessive memory load (Sweller et al., 2011). The element interactivity effect suggests that the level of intrinsic cognitive load associated with element interactivity is equally important as extraneous cognitive load in determining the likely effectiveness of any CLT-based redesign. Since intrinsic cognitive load does not change when the task and learners' expertise level remain constant, most cognitive load effects work through reducing extraneous cognitive load to avoid overload. However, the nature of element interactivity makes it possible to manipulate intrinsic cognitive load by changing the initial structure of tasks. The variability effect, gained by manipulating intrinsic cognitive load, is discussed in the next subsection.

2.4.6 The Variability Effect

In contrast to extraneous cognitive load that should always be decreased, intrinsic cognitive load should be optimised rather than decreased (Sweller et al., 2011). The variability effect is a CLT effect that works by increasing intrinsic cognitive load, thus promoting information processing. Specifically, the variability effect occurs when highly variable learning materials result in better learning performance compared to less variable materials (Sweller et al., 2011).

Within cognitive load theory, the variability effect was first investigated by Paas and Van Merriënboer (1994) in geometrical learning experiments, showing that students in the high-variability group had better transfer performance when applying the worked-example strategy. In Paas and Van Merriënboer's (1994) study, a 2 x 2 between-subjects group design (worked example-high variability, worked example-low variability, problem solving-high variability, problem solving-low variability) was applied. Students in all groups were required to learn six problems; the worked example groups presented the questions with solutions, while the problem-solving groups were only presented the questions to solve. As for the level of variability, the high variability groups and low variability groups were presented with three identical questions and three different questions. The first, third, and fifth questions were the same, and the second, fourth, six questions were the variants of the standard questions. In low variability groups, the variants only differed by changing the values of the number in the questions, while in high variability groups, both values and question formats were changed (Paas & Van Merriënboer, 1994). An interaction effect between the variability level and the instructional method was found in this study, suggesting that the high variability was more effective in the worked examples instruction but not in the problem-solving instruction.

The mechanism of increased variability supporting learning lies in its promotion of germane processing (Paas & Van Merriënboer, 1994; Sweller et al., 2011). With high variability learning materials, learners have the opportunity to engage in deeper processing and construct a schema with broader applicability (van Merriënboer & Sweller, 2005). Increased variability changes the nature of the task, resulting in an increase in intrinsic cognitive load (Sweller et al., 2011). As intrinsic cognitive load increases, the opportunity to allocate more working memory resources to intrinsic cognitive load emerges. Thus, germane processing increases accordingly. Specifically, when problem variability is low, learners only need to learn the specific way to solve the problems. The deep structure underlying the problems is easily ignored because all problems can be solved without considering the deep structure. However, as variability increases, learners must take into account more and more elements associated with the various structures reflected in the problems. Deeper processing is necessary to gain a more fundamental understanding to solve all problems. As a result, learners gain a deeper understanding and demonstrate better learning outcomes.

However, high variability learning materials do not necessarily lead to better learning performance. If the variability imposes too much intrinsic cognitive load on learners, causing their total cognitive load to exceed the limit, the learning will not be promoted, and may even be harmed. As Paas and Van Merriënboer (1994) hypothesised, high variability would have a positive effect on learning when initial cognitive load was low, because the total cognitive load remained within working memory capacity limits, irrespective of the fact that variability increased intrinsic cognitive load. On the other hand, the positive effect of the high variability may disappear when initial cognitive load is already high, because the total cognitive load will overload the capacity of working memory. Their experiments supported these hypotheses: the high variability materials were only effective in supporting transfer question problem-solving under the worked example condition (i.e., low initial cognitive load), rather than under the problem-solving condition (i.e., high initial cognitive load). To achieve variability's supportive effect, both of the following conditions must be met. First, intrinsic cognitive load needs to be increased so that more working memory resources can be allocated to processing it. Second, the total cognitive load must not exceed a student's working memory limit to ensure that there are sufficient working memory resources available to be allocated to process the intrinsic cognitive load. Only in this way can we ensure that the newly added intrinsic cognitive load resulting from increased variability is handled appropriately, thereby improving germane processing and promoting learning.

CLT researchers are neither the first nor the only ones interested in variability. The impact of variability on learning has attracted interest from researchers in a range of other fields, including linguistics, pedagogy, sports coaching, and other domains over the past 80 years (Raviv et al., 2022). Variability's effect on motor learning (e.g., Moxley, 1979), categorization (e.g., Quilici & Mayer, 1996; Vukatana et al., 2015; Wahlheim et al., 2012), visual perception (e.g., Clopper & Pisoni, 2004), language acquisition (e.g., Singh, 2008), machine learning (e.g., Gliozzi & Plunkett, 2019; Hill et al., 2020), and learning in educational settings informed by both the variation theory of learning (Marton & Booth, 1997) and CLT (e.g., Likourezos et al., 2019) has been demonstrated.

Though variability has been studied for many years in various fields, there is no unified definition of variability among those fields. To provide a clearer picture of variability and learning, Raviv et al. (2022) integrated results from different domains and identified four kinds of variability, including a) numerosity (set size), concerning learning from more or fewer different examples; b) heterogeneity (difference between examples), concerning whether examples are more or less similar to one another; c) situational (contextual) diversity, concerning environmental conditions unrelated to the specific topic of instruction that are more or less variable; and d) scheduling (e.g., interleaving, spacing), concerning more or less varied learning practice schedules. Take learning tennis as an example, comparing the effect of training tennis serves repeatedly from just one location (e.g., 6 inches to the right of the centre mark) versus from four different locations (e.g., 6, 7, 8, 9 inches to the right of the centre mark) would be testing the variability from numerosity. Though the heterogeneity is also constituted (6 inches is naturally different from 7, 8, and 9 inches), this is not a deliberately manipulated heterogeneity. As for heterogeneity of variability in research design, learning tennis from four locations that are quite close to each other (e.g., 6, 7, 8, 9 inches to the right of the centre mark) versus learning tennis from four locations that are spread apart (e.g., 6, 12, 18, and 24 inches to the right of the center mark) can better test variability from heterogeneity. To test variability from situational (contextual) diversity, the research design could compare the effect of learning tennis on a court painted with one colour only (e.g., green) versus on various courts painted with various colours (e.g., green, red, blue, orange, etc.). As for the last type of variability from scheduling, the best example is comparing the effect of learning tennis in blocked practice (e.g., 6, 6,6, 12, 12, and 12 inches to the right of the center mark) verses in interleaving practice (e.g., 6, 12, 6, 12, 6 and 12 inches to the right of the center mark). As Raviv et al. (2022) concluded, many studies may confound numerosity and heterogeneity and treat them as interchangeable concepts. Though manipulating the variability from numerosity would be highly likely to encompass heterogeneity, their fundamental focus is different. Apart from where these two types of variability are confused, most studies have focused on one single source of variation and developed results based on it (Raviv et al., 2022).

Notwithstanding the wide range of scholarly considerations of variability, there is a typical finding that more variability may first hinder learning but then benefit generalisation or transfer, indicating that the mechanism behind different kinds of variability may be somewhat similar. Raviv et al. (2022) reviewed three potential mechanisms behind hundreds of empirical studies. The first explanation is that variability can help learners to identify task-relevant dimensions and establish correct decision boundaries. Students can better notice the core part of tasks by showing variable examples with some constant elements. Variation theory (Marton, 2015; Marton & Booth, 1997), drawing on a pedagogical perspective, posits a similar argument that variation is a necessary component in teaching for students to notice what is to be learned. The second explanation is that variability may provide greater coverage of task-relevant space and boost learners' generalisation, extending from extrapolation to interpolation. The third explanation focuses on variability's linkage to retrieval performance, arguing that variability can promote retrieval performance through a more effortful cycle of forgetting and reconstruction. The desirable difficulty hypothesis (Bjork & Bjork, 2011) has a similar underlying philosophy to this explanation, arguing that retrieving or generating more information could enhance learning.

In summary, the variability effect works by manipulating intrinsic cognitive load and optimising it (Sweller et al., 2011). The supportive effect of variation on learning has been studied and demonstrated in various domains. CLT explains the variability effect from a fundamentally cognitive perspective, not only explaining the mechanism of why higher variability promotes learning but also clarifying why it only becomes effective under specific conditions. The significant interaction between variability level and other variables (e.g., instructional type, expertise level) has supported the argument that high variability only has effects when working memory capacity is sufficient to handle the increased cognitive load (Likourezos et al., 2019; Paas & Van Merriënboer, 1994), suggesting that variability has the potential to alter the characteristics of the simple tracing effect and better promote learning.

As more compound cognitive load effects have been found and developed in recent years, the theoretical framework of cognitive load itself has advanced compared to when it was first proposed. Evolutionary psychology, providing a strong foundation for human cognitive architecture, was incorporated into the CLT theoretical basis. Two fundamentally distinct categories of knowledge from an evolutionary perspective and their application to support learning (e.g., animation, human movements) are described in the next section.

2.5 Incorporation of Evolutionary Perspective into CLT

All human structures and functions, including physical and cognitive aspects, are products of evolution by natural selection. Compared to the awareness of evolution's impact on physical structure, research into the evolution of the human mind is less developed. However, considering human brain structures and cognitive functions from an evolutionary perspective, seeing them as the product of evolution by natural selection can trigger new insights into the mechanisms of the human mind, and further bring new insights into instructional applications (Sweller, 2003, 2004; Sweller et al., 2011; Sweller & Sweller, 2006).

This section firstly introduces five basic principles encompassed within the natural information processing system, which underlie both biological evolution and human cognition. Then two fundamentally distinct categories of knowledge human acquired from evolved, *biologically primary knowledge* and *biologically secondary knowledge*, are discussed in this section, followed by their implications for instructional design.

2.5.1 Five Principles Underlying Biological Evolution and Human Cognition

The process of biological evolution by natural selection can be understood as the process by which genetic information is stored and passed on to succeeding generations, with random mutations of the genetic material. If the mutations are adaptive to the environment, then the mutations are retained and stored in the genetic information, otherwise they are abandoned (Sweller et al., 2011). This information processing system is equally applicable to the processing of knowledge and information pertaining to human cognition. Both biological evolution and human cognition are sophisticated natural system that create, disseminate, use and store information, relying on the same basic information processing machinery to function (Sweller et al., 2011; Sweller & Sweller, 2006). The underlying machinery is based on five principles, explaining how information is acquired, processed, stored, and applied in various environments (Sweller, 2022; Sweller et al., 2011).

The first principle is *the information store principle.* To deal with the complexity of various natural environment, natural information processing systems store large amount of information (Sweller et al., 2011; Sweller & Sweller, 2006). All genomes consist of thousands, even billions of base pairs that could be considered as information units, carrying large amount of information that help species to face the inevitable environmental variations during the process of biological evolution (Sweller et al., 2011). The character of long-term memory in human cognition is equivalent to genomes in biological evolution, both of which provide very large information stores to support our daily life (Sweller, 2022; Sweller et al., 2011). With huge amount of information stored in long-term memory, human can engage in activities from automatically recognising objects to planning routine daily lives (Sweller et al., 2011). Information held in long-term memory is stored in the form of schemas, which are essential for human to function. The immense size of long-term memory store is sufficiently large to enable the variety of cognitive activities, constituting the most fundamental and central principle underlying the information processing machinery (Sweller et al., 2011).

The second principle is *borrowing and reorganising principle*. This principle explains one of the basic processes by which information is acquired: obtaining organised information from other information stores (Sweller et al., 2011). From the perspective of biological evolution, a genome is passed from one generation to the next, carrying the information stored in the preceding organism and passing it on to the next organism. In the case of asexual reproduction, a genome is exactly copied and "borrowed", while in the case of sexual reproduction, a genome is borrowed from both male and female ancestors and reorganised to form a unique genome for the next generation (Sweller et al., 2011). Human cognition acts in a similar way as biological evolution for information acquisition. Most information stored in one's long-term memory is borrowed by imitating other (Sweller & Sweller, 2006). Knowledge stored in long-term memory can be transmitted to others. In most situations, the borrowed knowledge is reorganized and processed through previously stored schemas in long-term memory, constructing a new schema that is different in informational content to the schema being borrowed (Sweller et al., 2011).

The third principle is the *randomness as genesis principle*. This principle introduces another process that natural information systems use to acquire information: creating novel information (Sweller et al., 2011). The borrowing and reorganizing principle only allows human acquire information from others, while the randomness as genesis principle explains the capacity for creativity (Sweller & Sweller, 2006). From the perspective of biological evolution, random mutations introduce new information into a genome and cause genetic variation between individual organisms. Mutations adaptive for survival and reproduction are retained in the species genome, while those that are not effective for survival are eliminated (Sweller et al., 2011). As for human cognition, similar "mutation" also occurs. When a person attempts to solve a novel problem where relevant knowledge is absent, random attempts of possible moves will be made (Sweller et al., 2011). The process of these random new attempts is analogous to mutation, introducing new information (new solutions to novel problem) to human cognition. Just as only a proportion of naturally mutated genes are retained, constituting evolution, only the tested solutions that are effective are stored in long-term memory, constituting learning (Sweller & Sweller, 2006). In summary, the information stored in long-term memory is either acquired by borrowing, reorganizing, or creating.

As the ultimate purpose of acquiring and storing huge amount of information is survival in a complex external environment, the following two principles focus on the required machinery (Sweller et al., 2011), further explaining how natural information processing systems interact with the environment.

The fourth principle is the *narrow limits of change principle*. This principle explains that creation of new information is naturally limited, in order to avoid potentially negative consequences of rapid random changes (Sweller & Sweller, 2006). In the case of biological evolution, the epigenetic system acts as an intermediary between the genetic system and the external environment (Sweller et al., 2011; Sweller & Sweller, 2006), mediating the survival stress from the environment and organismal functions rooted in the genome. The epigenetic system selectively processes and transmits information from the external environment to the genetic system in a manner that can result in genetic changes. Mutation in some parts of a genome can be facilitated while mutations in other parts of a genome can be inhibited, depending on the adaptability of the epigenetic content exhibited by that part of genome to the external environment (Sweller et al., 2011). As the content of mutation is random, and mutations need to be tested in small steps for their adaptivity, changes in the genome are naturally limited. In the case of human cognition, working memory plays a similar role as the epigenetic system. Working memory act as an intermediary between long-term memory and the external environment, dealing with novel information from the environment and stored schema in long-term memory. With limits in capacity (e.g., Cowan, 2001; Miller, 1956) and duration (Peterson & Peterson, 1959), working memory can only process a small amount of information simultaneously, limiting the changes in long-term memory.

The fifth principle is the *environmental organising and linking principle*, explaining how huge amounts of stored information can be used to determine activity relevant to a particular environment. In the case of biological evolution, the epigenetic system controls the link between genetic characteristics to physical characteristics. As introduced in the narrow limits of change principle, the epigenetic system connects the genome and environment via determining where mutations are facilitated or inhabited, while under the environmental organising and linking principle, the epigenetic system connects the genome and environment by determining which stored genetic information will be used or ignored (Sweller et al., 2011). Compared to the limit in genome changes, the amount of genetic information required to express particular physical characteristics has no limit. This is because huge amounts of genetic information has already been stored in the genome and the epigenetic system can organise this information to determine a particular physical character (Sweller et al., 2011). As in the case of human cognition, working memory plays a similar role. As introduced in the narrow limits of change principle, working memory can receive novel information from the external environment and store it in long-term memory. Working memory also can interact with stored information in the long-ter memory, using it to coordinate activity with the environment (Sweller et al., 2011). Similarly, when dealing with organised information stored in long-term memory rather than random information generated from the external environment, information processing is not constrained by the working memory limits on capacity and duration.

In summary, the five basic principles of natural information processing systems are equivalently applied to biological evolution and human cognition (Sweller & Sweller, 2006). The first four principles explain how information is stored, acquired, created, and retained, while the fifth principle explains how stored information is applied to face environmental requirement. However, it should be noted that these principles are mostly relevant for biologically secondary knowledge, not biologically primary knowledge. Detailed aspects of these two types of knowledge are introduced in the next section.

2.5.2 Biologically Primary and Secondary Knowledge

From an evolutionary perspective, Geary (2008) classified all knowledge into biologically primary knowledge and biologically secondary knowledge. As the name suggests, biologically primary knowledge is the type of knowledge that is most fundamental for biological survival. In contrast, the biologically secondary knowledge is the type of knowledge that is required for cultural reasons. To distinguish these two categories of knowledge, Geary (2008) suggested that biologically secondary knowledge is learnable and teachable, while biologically primary knowledge is learnable but not teachable because biologically primary knowledge is usually learned without being deliberately taught.

According to Geary's (2008) theory, humans have evolved to assimilate biologically primary knowledge. It can be seen as an instinct programmed into the human genome that acquires biologically primary knowledge without being taught (Sweller et al., 2011). Biologically primary knowledge constitutes a wide range of basic skills that human beings need to survive in the world. Some skills are universal, basic, and acquired effortlessly, so that we often take them for granted (Sweller et al., 2011). For instance, recognising and distinguishing human faces is biologically primary knowledge, with this capability first demonstrated by human beings as early as the first several days of birth (Walton et al., 1992). Considering how much detail a face can contain, it is impossible to teach a one-day-old infant to recognise this particular face by teaching the infant to memorise features like the nose, eyes, and mouth. In fact, infants can distinguish between different faces. Therefore, recognising the human face is a skill that can be acquired but cannot be deliberately taught. Humans evolved to easily gain this biologically primary knowledge because the ability to identify conspecifics and social agents is critical for our survival (Simion & Giorgio, 2015). As a biologically primary skill, the figures of nose, eyes and mouth are automatically processed as an integrated face, and made available to consciousness, without requiring effortful processing in working memory (Geary, 2008). Besides face recognition, many other basic skills like speaking the first language, and physically interacting with the environment through movement, are also based on biologically primary knowledge.

Biologically primary knowledge is considered modular, with each biologically primary skill being largely independent of other skills (for a taxonomy of the proposed biologically primary modules, see Figure 1 of Geary, 2008). Each biologically primary skill likely evolved separately from other primary skills and had its own evolutionary path over the long history of human beings, which enabled each biologically primary knowledge to have its own acquisition system (Sweller et al., 2011). For example, research in neuroscience has found that the human brain is equipped with a neural circuitry specialised for preferentially processing faces (Haxby et al., 2002). Because of evolutionary history, various biologically primary skills can be acquired easily and effortlessly at a very young age without explicitly being taught (Sweller et al., 2011).

Biologically secondary knowledge, on the other hand, is a kind of “new” knowledge compared to primary knowledge. Since human culture is constantly evolving and our knowledge base is growing exponentially, there is a great deal of knowledge that has existed for far too short a time to be influenced by biological evolution (Sweller et al., 2011). The knowledge that human beings were not evolved to gain is referred as biologically secondary knowledge, which is culturally important for us to obtain (Paas & Sweller, 2012; Sweller, 2022). Acquiring biologically secondary knowledge does not confer a direct advantage for survival and reproduction, which prevents such knowledge from being preserved in the human species' genes through natural selection (Sweller & Sweller, 2006). For instance, speaking the first language is a biologically primary skill because communication with others is a fundamental skill that enables humans, as social creatures, to survive. In contrast, writing a first language is a biologically secondary skill, because the inability to write does not prevent people of this era from surviving and reproducing in this world. For the vast majority of the time after humans learned to write, only a tiny fraction of the population possessed the skill of writing. This did not prevent those who could not write from living their lives. Though we may require many aspects of biologically primary knowledge in order to learn to write, we have not evolved to learn to write (Sweller et al., 2011). Besides writing, every curriculum area from practical fields (e.g., cooking) to academic fields (e.g., mathematics) that requires instructional support consists heavily of biologically secondary knowledge.

In contrast to biologically primary knowledge having their own acquisition systems, humans have not evolved any system to acquire any particular biologically secondary skill. However, humans have evolved a cognitive system that permits the acquisition of a possibly infinite range of secondary skills (Sweller et al., 2011). Unlike biologically primary knowledge based on multiple separate modules, biologically secondary knowledge is argued to have a unitary system to acquire all biologically secondary knowledge (Geary, 2008). As the generation of new knowledge is a rapid and continues process, whereas the biological evolution is a remarkably slow process, it is impossible for humans to be born with particular processing systems to address every type of knowledge. A general information processing system is necessary to deal with all the information that cannot be processed through any particular biologically primary system. When a person encounters biologically secondary information, which cannot be automatically processed through any biologically primary acquisition system, attention will be automatically shifted to this information, and the information will be stored in working memory awaiting processing. This ability to represent and manipulate information in the working memory was is the core component of the human ability to adapt to ecological and social variation within a life span (Geary, 2007).

In summary, the fundamental distinction between biologically primary and secondary knowledge lies in whether they are basic and essential to survive, and based on this, whether humans have evolved to acquire such knowledge. This fundamental distinction gives biologically primary knowledge and biologically secondary knowledge different characteristics. Biologically primary knowledge is learnable but not teachable, while biologically secondary knowledge requires explicit teaching. Biologically primary knowledge is learned easily, automatically, and unconsciously in a functioning social context (e.g., families and broader communities), while biologically secondary knowledge is learned consciously and requires effort. These characteristics reveal that biologically secondary knowledge cannot be acquired via immersion (Sweller, 2008), suggesting that most instructional design applied in educational institutions should focus on biologically secondary knowledge.

Since biologically primary knowledge is effortless to acquire, Paas and Sweller (2012) proposed that the biologically primary system could be used to facilitate learning biologically secondary knowledge. Though previous arguments suggested that teaching biologically primary knowledge may be futile, those biologically primary, genetic-cognitive skills can be used to assist in teaching biologically secondary, domain-specific skills (Paas & Sweller, 2012). Showing students how a biologically primary skill can be used in a specific course or field can be instructionally effective (Sweller et al., 2019; Youssef-Shalala et al., 2014). As Sweller et al. (2019) summarised, teaching anything involves a combination of primary and secondary skills, while the secondary skill is the only part that needs to be learned.

Biologically primary knowledge is also argued to assist the acquisition of biologically secondary knowledge via intrinsic motivation. Learning biologically primary knowledge is driven by strong intrinsic motivation because it is closely related to survival (Geary, 2008), while the specific actions of learning biologically secondary knowledge, such as doing homework and listening to lectures, were reported as the lowest level of happiness in everyday life (Csikszentmihalyi & Hunter, 2003). Lespiau and Tricot (2022) found that incorporating biologically primary knowledge into learning a biologically secondary topic (logical syllogisms) not only enhanced participants' motivational beliefs but also encouraged participants to be motivated throughout the task. This research demonstrated how biologically primary knowledge benefited biologically secondary knowledge, suggesting that biologically primary knowledge should not be put aside simply because it has already been learned

(Lespiau & Tricot, 2022).

More specifically, a range of particular types of biologically primary knowledge have been found to facilitate biologically secondary knowledge learning, including communication (the collective working memory effect), understanding animation (human movement effect), and sensorimotor experiences (embodied cognition) (Paas & Sweller, 2012). Embodied cognition, specialised in sensorimotor experiences like gesture and pointing, is discussed in the next section.

2.5.3 Embodied Cognition, Gesture, and Pointing

Embodied cognition emerged during the natural development of cognitive science in the late 20th century. Barsalou (1999) proposed a model of the perceptual symbol system, arguing that there are two different kinds of symbols, amodal and modal symbols, in our cognitive system. This argument opposes Fodor's (1975) standpoint that abstract symbolic representations make up humans' conceptual system. In Barsalou's (1999) description, the amodal symbols (abstract symbols) do not contain perceptual contents, while the modal symbols (perceptual symbols) are in the same neural systems that are used when the referents are initially perceived. This argument for the same neural systems for perception, action, and mental simulation of an entity explains why symbols used in cognitive processes also contain perceptual contents, giving the theory base that sensory factors may affect cognitive learning.

The embodied cognition perspective assumes that all cognitive processes are based on action and perception instead of abstract constructs (Barsalou, 1999), supporting cognition through goal-directed interactions between organisms and the environment (Barsalou, 2008; Glenberg, 1997; Rueschemeyer et al., 2009). As Paas and Sweller (2012) explained, "Cognitive representations of symbols like numbers and letters are ultimately based on sensorimotor codes within a generalised system that was originally developed to control an organism's motor behaviour and perceive the world around it, which has resulted in automaticity of perceptual and motor resonance mechanisms in cognitive tasks" (p35-36). Research from various domains have provided evidence for the embodied cognition framework. Specifically, neurophysiological research has found the brain's motor system was active during the comprehension of language (Glenberg et al., 2008); research from neuroimaging has demonstrated that information about salient properties of an object is stored in the sensory and motor systems (Martin, 2007); and research from semantics has found that semantic presentation may trigger or resonate with motor action (Lindemann et al., 2006; Zwaan & Taylor, 2006). Taken together, these lines of research suggest that cognitive and sensorimotor processes are closely intertwined (Paas & Sweller, 2012).

From evolutionary educational psychology, gesture, which humans start to make as young as 12 months of age (Liszkowski et al., 2012), can be seen as a typical sensorimotor experience that specifically demonstrates how biologically primary knowledge assists in biologically secondary knowledge's acquisition through embodied cognition (Paas & Sweller, 2012). In the education field, gestures are considered particularly important in the interaction between students and teachers. Gestures from the students help the teacher better understand students' learning processes and progress (e.g., Alibali & Goldin-Meadow, 1993), and gestures from teachers give students more information on instruction (e.g., Flevares & Perry, 2001). It was found that learners understood the teacher's instruction that combining gesture and speech better than the instruction with speech or gesture only (Kelly, 2001).

Various studies demonstrated that both observing gestures (e.g., Church et al., 2004; Perry et al., 1995) and making gestures (e.g., Broaders et al., 2007; Cook et al., 2008) can improve learning outcomes. Studies also found that giving verbal explanations with gestures can reduce cognitive load during instruction in mathematics (Goldin-Meadow et al., 2001) and Piagetian conservation tasks (Ping & Goldin-Meadow, 2010). Goldin-Meadow and colleagues (2001; 2010) have argued that gesture can convey information through a visuospatial format, in contrast to speech transmitting information through verbal representations. As such, gesture has the potential to enrich the way information is encoded, open a new channel for working memory, and expand working memory's capacity, thus promoting knowledge acquisition.

As particular types of actions on objects (Congdon et al., 2018), *pointing* (touching a surface with a finger in a static position) and *tracing* (touching a surface with a finger in an active movement) have some similar features to gestural hand movements. Both of them contain motor perception in their process, which can also be explained through the embodied cognition framework as a support for cognition and learning. Pointing was found to support learning in different domains, including language learning (e.g., LeBarton et al., 2015) and mathematics learning (e.g., Alibali & DiRusso, 1999). Zhang et al. (2023) focused on finger pointing as a self-management approach to control cognitive load, finding that it led to higher retention performance than no pointing. Comparing with pointing via a computer mouse, finger pointing demonstrated its superiority in facilitating learning (Zhang et al., 2023), which further supports the argument from the evolutionary perspective that finger pointing worked more as biologically primary knowledge, rather than merely an indicator, to facilitate cognitive learning.

In recent years, research has applied the pointing and tracing strategy together, asking students to use their hands to understand learning materials via pointing and tracing, and has found that students in this group performed better than the control group (e.g., Ginns & King, 2021; Ginns & Kydd, 2019; Macken & Ginns, 2014). In addition to activating the proprioceptive system through movement, tracing also generates tactile sensation, which may facilitate learning and lower cognitive load to a greater extent than gesture alone (Hu et al., 2015). Tracing, as a more complex and varied gesture and the key factor in this PhD study, is discussed in the next chapter.

Chapter Three: Tracing

As biologically primary skills, index finger movements (pointing and tracing) evolved to be acquired by humans easily at a very young age. Infants are found to use index finger pointing to express their needs before one year old (Boundy et al., 2019; Liszkowski et al., 2004). The capacity for finger tracing, as dynamic finger movements distinct from static finger pointing (McNeill, 1992), is also coded within the human genome. Archaeological research indicates finger flutings (moving fingers on soft surfaces and thus leaving a mark) are associated with cave art back to 50,000 years ago (Walshe et al., 2024). Apart from the long history, the widespread geographical distribution (across Europe and Australia) of finger flutings also suggests that this activity existed in the early stages of human evolution (Bednarik, 1986). Marquet et al. (2023) recently found that the Neanderthals, close relatives of human beings, also made such finger flutings, and they demonstrated that such finger flutings were made from intentional finger tracing actions (which they called "engraving"), rather than accidental drawing. Children, as young as 2 years old, were suggested to be the possible creators of some finger flutings (Sharpe & Van Gelder, 2006; Van Gelder, 2015). Taken together, these archaeological research findings indicate that finger tracing is a historical and culturally universal human activity that can be acquired in early childhood, suggesting that finger tracing is a biologically primary skill that humans evolved to acquire.

The history of co-opting finger tracing into instruction to support learning originated over a century ago when *Sandpaper Letters* were invented by Maria Montessori (see Montessori, 1912). Sandpaper letters are a set of teaching materials and the corresponding instructional method used to teach children letters. The letters of the alphabet were cut out from the sandpaper and glued onto smooth cards. In this learning activity, children trace letters cut out of sandpaper with their fingers in the same sequence as writing the letter, with the teacher pronouncing the letter's sound at the same time (Montessori, 1912). Various sensations, including visual, auditory, tactile, and muscular sensations, work collaboratively to promote learning. From letters to words, Montessori incorporated sandpaper letters into a coherent teaching programme, enabling four-year-old children to compare with primary school students after six months of learning (Montessori, 1912). In recent years, empirical studies have demonstrated that finger tracing can enhance learning, which is discussed in the following sections.

3.1 Empirical Studies on Tracing and Learning

Originating from sandpaper letters, finger tracing then became a widely used method in the field of education, operationalised as moving the index finger in an established trajectory to trace the key information in the learning material. As a learning strategy that is easy to implement but effectively promotes learning, finger tracing attracted the attention of educational researchers. The focus of tracing extends beyond academic performance, with attention also being paid to non-academic (emotional) aspects. Empirical studies on tracing's effect on learning performance, intrinsic motivation, and cognitive load are separately discussed in the following subsections.

3.1.1 Effects on Learning Performance

Empirical findings on letter learning supported Montessori's idea that finger tracing could enhance learning. Hulme (1981, Experiment 1) compared the visual method and the visual plus tracing method for letter learning in 10-year-old normal and developmentally delayed readers. Children were displayed with a set of 14 letters and asked to point and read (visual condition, V), or trace and read (visual plus tracing, VT) the letter. Results showed that children in the VT condition remembered more letters than those in the V condition. An interaction effect was also found between condition and reading ability, suggesting that tracing proves more beneficial for delayed readers. Hulme et al. (1987) then extended the experiments to a broader age range and non-letter learning, demonstrating that tracing's effect could be found with preschool children (Experiment 1) and with abstract letter-like forms (Experiment 2 to 4), suggesting that tracing facilitated letter learning because it helped learners to recognise visual forms. Bara et al. (2004) conducted similar research in France and found that incorporating tracing in letter learning sessions could help children more easily connect letters' orthographic representation and corresponding sounds. This study compared VAM (visual-auditory-metaphonological) training, in which children were asked to follow the drawing of the letter with their eyes, and HAVM (haptic-visual-auditory-metaphonological) training, in which children were asked to run the index finger along the letter's outline. Results showed that the HAVM group performed better on the pseudo-word decoding task. Bara et al. (2007) further found that HAVM training not only improved letter recognition and initial phoneme identification performance in the post-test, but also improved pseudo-word decoding performance in a delayed post-test conducted more than six months later.

The role of finger tracing extends far beyond letter learning. Empirical studies also demonstrated its effect in similar fields, from abstract letter-like shapes recognition (Hulme, 1979, 1981, Experiment 2 and 3) to further word reading (Jensen & King, 1970). Jensen and

King (1970) compared three different learning strategies on preschool children's word learning: tracing the stimulus word in textured form, rearranging constituent letters to form the stimulus word, and choosing the word that matches the stimulus word. Within each strategy, children were further divided into two groups depending on whether they learned the similar words set or the dissimilar words set. The tracing group performed better than the matching group and the rearranging group in the training session, but no difference was found between the three strategies in the post-reading test performance.

In addition to abstract letter-like shapes, tracing has also been found to support the learning of geometric shapes. Kalenine et al. (2011) taught preschool children three categories of plane geometrical shapes (square, rectangle, and triangle) via two different interventions: multisensory intervention (VH condition, visual plus haptic) and classic intervention (V condition, visual only). Children in the VH condition progressed more in their recognition of rectangles and triangles compared to those in the V condition, but no difference was found on squares. The benefits of tracing for learning geometry are not confined to simple recognition learning; it is also reflected in more advanced topics, such as geometric rules. For example, F.-T. Hu et al. (2014) tested tracing in the learning of angle relationships involving parallel lines, and demonstrated tracing's learning supportive effect in geometric rules learning. In this research, Year 6 students from Australia were randomly allocated to the tracing and non-tracing conditions. Students in the tracing condition were asked to read the worked example and trace the vertical angles, while students in the non-tracing condition were asked to read the worked example with their hands placed on their laps. Results showed that students in the tracing condition solved more practice problems in the acquisition phase and made fewer errors in the test phase. A ceiling effect was found on test questions correctly answered, where 78.6% students correctly solved all questions. F.-T. Hu et al. (2015) later conducted two similar experiments on Year 5 students. Experiment 1 addressed shortcoming of ceiling effect in previous study and demonstrated tracing effect again. They recruited younger students (Year 5) with less mathematics experience to avoid ceiling effects and added a new indicator “time to solution” to measure learning performance. In Experiment 2, F.-T. Hu et al. (2015) compared three conditions (no tracing vs. tracing above the paper vs. tracing on the paper), aiming to clarify whether tactile and kinaesthetic modes individually affect learning. Results showed that tracing on the paper group had best performance, following by tracing above the paper group, and the no tracing group. This experiment suggested that both tactile and kinaesthetic modes engaged in enhancing learning. F.-T. Hu et al.’s (2014, 2015) study was then translated and replicated by other researchers in China (Du & Zhang, 2019) and Malaysia (Yeo, 2024; Yeo & Tzeng, 2020), demonstrating tracing effect on geometry rules in different language and culture backgrounds.

Other than geometry learning, tracing has been found to be applicable to a wider range of mathematics learning. Agostinho et al. (2015) demonstrated that finger tracing could better support mathematical graphs (temperature graphs) learning, helping students perform better in the transfer questions. Ginns et al. (2016, Experiment 2) further investigated tracing’s effect on learning order of operations, a topic that does not involve diagrammatic elements. In this experiment, students in the tracing condition were asked to trace the operation symbols in the worked examples, while students in the non-tracing condition were only asked to read the worked example. Students in the tracing condition had marginally better performance in the practice questions and significantly better performance in the post-test. This research demonstrated that finger tracing could not only support complicated learning in graphics-related fields, but also support learning in topic areas that are less inherently spatial.

Besides the letter, word, shape, geometry and mathematical learning mentioned in the previous paragraph, the tracing effect has also been examined in the context of learning from expository text and diagrams. Macken and Ginns (2014) tested tracing with the learning materials about the structure and function of the human heart, and found that learners who applied pointing and tracing performed better on tests of terminology and comprehension, demonstrating tracing's effect in the science and medicine domain. This study on tracing effect on human physiology was then replicated by Ginns and Kydd (2019), finding that students in the pointing and tracing group had better performance in both terminology test and comprehension test. Y. Wang et al. (2025) translated the learning material and replicated earlier findings with Chinese native speakers, finding students in the pointing and tracing group had better performance on a comprehension test. Tang et al. (2019) investigated tracing on the water cycle lesson materials and found that students in the tracing group had better performance in on recall and transfer tests. Taken together, these studies showed that finger tracing is a useful strategy to enhance learning in various learning domains.

The role of tracing in promoting learning is demonstrated not only in recognition and retention tasks, but also in transfer tests. Earlier tracing studies on letter learning demonstrated that tracing could help children better recognise the letter or abstract shapes (e.g., Hulme, 1981; Hulme et al., 1987). Specifically, the researcher asked children to name the letter or abstract item after learning and evaluated whether they correctly recognised them. More recent tracing research tends to consider both acquisition and transfer simultaneously. For instance, Agostinho et al. (2015) tested students' learning performance via similar questions and transfer questions. Among 10 test questions in their experiment, three questions were defined as similar questions because they had the same structure and required the same procedural steps to solve as presented in the acquisition phase. The remaining seven questions were transfer questions because they required students to transfer their understanding of the learning to solve. Results showed that the tracing group performed better on both similar questions and transfer questions, indicating that finger tracing not only helps students better memorise the steps for answering questions, but also aids them in better understanding the learning material. However, some studies only detected tracing effect on transfer questions, which may be attributable to the ceiling effect on similar questions (e.g., Ginns et al., 2016). Du and Zhang (2019) further distinguished near transfer and far transfer questions, finding that the tracing effect sometimes only occurs on far transfer questions (Experiment 2).

Though no definitive conclusion has been reached at present, tracing seems to yield better effects in more complex transfer questions. Transfer refers to the ability to use what has been learned to solve new problems, reflecting the extent to which learners have understood the learning content (Mayer, 1996, 2002; Mayer & Wittrock, 1996). Cooper and Sweller (1987) argued that the differential performance in similar questions and transfer questions is due to differential applicability of schemas. When answering test questions, learners can apply an acquired schema to answer similar questions, because the similar questions have the identical schema as the worked example. However, learners cannot directly use the schema to answer transfer questions, because transfer questions share the same rules as the worked example, but not the identical schema. Thus, rule automation is the primary factor that facilitated solution of transfer problems. In the context of finger tracing, tracing effects may be stronger on transfer questions than similar questions because finger tracing facilitated not only schema acquisition but also rule automation.

Regarding tracing's effectiveness across age groups, finger tracing has been demonstrated to be effective across all age groups from early childhood to adulthood. As introduced in the previous paragraphs, many letter learning and recognition studies involve young children as participants, from three-year-old children (e.g., Hulme et al., 1987, Experiment 1) to five-year-old children (e.g., Kalenine et al., 2011). Finger tracing has also been demonstrated as a helpful method for school-aged children in their subjects' learning, including primary school students (e.g., Agostinho et al., 2015, Grade 3) and secondary school students (e.g., Ginns et al., 2016, Experiment 1, Grade 7). As for adult learning, numerous empirical studies have demonstrated that finger tracing supports learning among college students (e.g., Ginns & King, 2021; Ginns & Kydd, 2019; Macken & Ginns, 2014; Wang et al., 2025). As a biologically primary skill that human gained at very young age, we instinctively use it to aid our learning. For instance, young children tend to use their finger gently follow the text while reading. Though such behaviours gradually diminish in adulthood, the supportive role of finger tracing in learning persists. This reduction in such spontaneous behaviour may be attributable to a decrease in activities requiring substantial cognitive resources in daily life as individuals age. Various empirical studies have demonstrated that finger tracing can still enhance learning when adult students engage in learning novel topics.

As a simple and practical learning strategy, finger tracing was found to be applicable across various teaching scenarios and can be integrated with multiple instructional designs. Tracing's effect has been demonstrated not only in one-to-one experimental intervention settings (e.g., Hu et al., 2014), but also in general classroom environments (e.g., Du & Zhang, 2019). The universality of tracing effects is also reflected in the learning materials. Earlier tracing studies followed Montessori's approach, using textured card or other convex material to enable students to trace its outline (e.g., Jensen & King, 1970). As learning content is no longer confined to simple letters, flat paper-based learning materials are increasingly being adopted (e.g., Hu et al., 2015). More recently, tracing on an electronic screen has also been demonstrated to enhance learning (e.g., Agostinho et al., 2015), suggesting that finger tracing can be effective on a wide variety of object surfaces. Regarding the instructional design, finger tracing can be employed across various pedagogical methods, including worked-example (e.g., Ginns et al., 2016), expository text with diagrams (e.g., Macken & Ginns, 2014), and illustrated poster (e.g., Tang et al., 2019).

In summary, finger tracing is a learning strategy that has a long history, and it has been extensively studied in recent years. Finger tracing was found to enhance learning in various domains, helping students to better master the knowledge, especially in the transfer questions. Learners from multiple age groups can benefit from finger tracing, from preschool kids to adults in college. Finger tracing was found to be widely used in multiple learning contexts, indicating that it was a highly practical learning strategy. The effect of finger tracing extends beyond enhancing learning performance; it was also found to affect learners' motivation and cognitive load.

3.1.2 Effects on Motivation and Cognitive Load

Since Montessori first invented sandpaper letters, tracing has been regarded as a method capable of influencing students' emotional factors. As she described, "children are particularly enthusiastic about tracing the sandpaper letters" (Montessori, 1912, p. 293). Geary (2008) considered learning and motivation from the perspective of evolutionary educational psychology, arguing that humans have evolved motivational biases that focus on learning in biologically primary domains. Specific to tracing, moving fingers as a biologically primary activity may enable students to experience higher levels of motivation. The following tracing studies are also interested in the impact of finger tracing on learners' motivation. Ginns and King (2021) found that university students in the pointing and tracing group reported higher intrinsic motivation in the computer-based astronomy course learning. Wang et al. (2022) applied two types of tracing in their experiments, and found similar results on school-aged children that both the tracing/tracing and the tracing/imagination group reported higher intrinsic motivation than the control group. However, this finding was not replicated in their subsequent studies on college students, in which all three groups reported similar levels of intrinsic motivation. Park et al. (2023) further distinguished three types of learning motivation in their study, separately reflecting the perspective of future goals, mastery, and interest in the motivation. However, no difference between the pointing, tracing and control group was found on any type of motivation.

Researchers' focus on tracing impacts is also reflected in the cognitive load perceived by learners. Drawing on Geary's (2008) distinction between biologically primary and secondary knowledge and incorporating the CLT framework proposed by Paas and Sweller (2012), many tracing studies have found that finger tracing helps learners reduce cognitive load.

Some studies considered general cognitive load or singular cognitive load. F.-T. Hu et al. (2015) designed the intrinsic cognitive load measurement based on test question difficulty ratings, to it easier for young children to understand. In Experiment 1, they found that the tracing group reported a marginally lower difficulty level on basic test questions and a significantly lower difficulty level on advanced test questions. In Experiment 2, they added another condition, tracing above the paper, to the study. Though no significant difference in test difficulty was found on basic questions, the results on advanced questions support their hypothesis, indicating that tracing on the paper group reported the lowest test difficulty, followed by tracing above the paper group, and the control group. A similar design that measures general cognitive load by task difficulty was applied in many studies (e.g., Agostinho et al., 2015; Du & Zhang, 2019; Ginns et al., 2016); however, their findings on tracing's effect on cognitive load remain diverse. Du and Zhang (2019) had similar findings as F.-T. Hu et al. (2015), indicating that finger tracing helps learners reduce cognitive load on more complicated questions. The finger tracing group reported the lowest task difficulty on far transfer questions, but similar task difficulty on near transfer questions in both experiments (Du & Zhang, 2019). However, Agostinho et al. (2015) found there was no difference in task difficulty reported by tracing and non-tracing conditions on both similar questions and transfer questions. Ginns et al. (2016) also found that tracing and non-tracing groups reported similar levels of task difficulty on all kinds of questions, which raised researchers' doubts about task difficulty rating as a measure of cognitive load. Though the idea applying task difficulty to measure intrinsic cognitive load originates from the theory that intrinsic cognitive load related to the complexity of learning content, it was argued that learners was found to have problems differentiating between intrinsic and extraneous cognitive load (Krieglstein et al., 2023), which may explain why some studies cannot find differences on cognitive load between experimental and control conditions.

Some other studies measured specific types of cognitive load to achieve a more sensitive result. Macken and Ginns (2014) separately measured intrinsic, extraneous, and germane cognitive load in their experiment. They referred to Cierniak et al.'s (2009) instrument and measured each type of cognitive load via one question. Contrary to their hypotheses, no difference was found in any type of cognitive load between the pointing and tracing group and the control group, which may be because the single-item measurement was not sensitive enough. Subsequently, researchers started to use the multi-item scale to gain a more accurate result. Tang et al. (2019) separately measured intrinsic (three items) and extraneous (three items) cognitive load in their experiments on school-aged students, referring to Leppink et al.'s (2013) instruments. They hypothesised that participants in the tracing group would report lower intrinsic and extraneous cognitive load than participants in the non-tracing group. However, only the hypothesis on extraneous cognitive load was supported by the results, as reflected by significantly lower extraneous cognitive load reported by the tracing group. No difference was found in intrinsic cognitive load between the two groups. Ginns and King (2021) found similar results on college students' learning that finger tracing reduced extraneous cognitive load, but not intrinsic cognitive load. They also used multi-item self-reports to measure cognitive load, referred to Leppink and van den Heuvel's (2015) instrument, with four items for intrinsic and four items for extraneous cognitive load. However, despite employing precisely the same measuring instruments, Ginns et al. (2020) found no difference between tracing and non-tracing conditions on both intrinsic and extraneous cognitive load in their experiment with school-aged children learning geometric rules from worked examples.

In summary, finger tracing's impact on motivation and cognitive load varied substantially across studies. Some studies have demonstrated finger tracing improved motivation and decreased cognitive load, which aligns with the cognitive load theory, while others have failed to find such differences between conditions, which may require further research and more sensitive measurements to explain. As tracing's effect on learning outcomes has been demonstrated by various studies, the underlying mechanism of its effectiveness has also been extensively discussed. Three principal theoretical perspectives on tracing effect are discussed in the next section.

3.2 Potential Mechanisms of the Tracing Effect

Tracing, as a biologically primary knowledge, has evolved to be learned by human beings (Geary, 2007, 2008), and can effectively support the acquisition of biologically secondary knowledge (Paas & Sweller, 2012). Regarding previous empirical findings, researchers have proposed several specific mechanisms through which tracing might support learning, including perspectives from haptics, attention, and cognitive load (for a review, see Park et al., 2019). Drawing on this review, the following subsections review potential mechanisms of the tracing effect, focusing on an expanded haptic working memory channel, enhanced attention near hand, and supported information organising and processing procedures.

3.2.1 Haptics is Another Working Memory Channel

As discussed in the previous chapter (see the modality effect), presenting learning information across multiple modalities (i.e., visual and auditory) can promote learning by making available additional working memory resources for learning. A similar mechanism can be argued for with haptics, where application of the tactile or kinaesthetic modality can activate the haptic channel in working memory (Kaas et al., 2008; Morimoto, 2020). In terms of finger tracing, this learning activity contains both tactile sensation and kinaesthetic trajectory sensation.

F.-T. Hu et al. (2015, Experiment 2) compared two different types of finger tracing (tracing on the paper vs. tracing over the paper) and a no-tracing condition. In two tracing conditions, participants were requested to do the identical tracing actions, except for whether their fingers touched the paper surface. Through this design, the tactile sensation has been effectively isolated from the tracing actions. Results showed that tracing on the paper condition had better performance on acquisition phase questions, test phase basic questions, and advanced questions than tracing over the paper condition; and both tracing conditions had better performance than the no tracing conditions. This experiment effectively demonstrated that the mere distinction of whether or not the learner touched the paper can yield differences in learning outcomes, thereby illustrating the importance of tactile sensation.

As for kinaesthetic sensation, various studies compared handwriting with typing, and found that handwriting could help learners to generate a correct sensorimotor representation and thus facilitate learning (Longcamp et al., 2005; Mayer et al., 2020). Compared to visual presentations that directly display the entirety of learning materials, the kinaesthetic sensation during finger tracing additionally encompasses the sequence of key learning elements' emergence, thereby providing an additional dimension of learning information. F.-T. Hu et al.'s (2015, Experiment 2) results, which showed that tracing over the paper condition performed better than the no tracing condition, also demonstrated that isolated kinaesthetic sensation could enhance learning.

Therefore, additional haptic sensory and haptic working memory generated by tracing is considered as one of the key mechanisms by which tracing can support learning.

3.2.2 Attention is Enhanced Near the Hand

The second possible explanation for tracing's effect on learning is the hand proximity effect. It describes a phenomenon in which differences exist in perceptual and attentional mechanisms for stimuli presented within the graspable space of the hand (Thomas & Sunny, 2019). A range of studies has demonstrated that humans naturally pay more attention to information near the hands (e.g., Abrams et al., 2008; Reed et al., 2006).

Based on this finding, a finger can act as a guide to concentrate the learner's attention in the tracing condition (Cosman & Vecera, 2010; Reed et al., 2006), supporting attentional focus on the most important information to learn. The application of eye-tracking technology provided firm evidence that pointing and tracing guided visual attention during learning (Korbach et al., 2020). Park et al.'s (2023) follow-up study conducted more detailed comparisons between the pointing, tracing, and control groups, finding that participants in the pointing group had a marginally longer fixation duration on the illustration areas. Park et al. suggested tracing actions may interfere with the visual presence of information in the moment of moving the finger, which renders tracing less effective than pointing.

To isolate the effects of attention, Du and Zhang (2019, Experiment 1) compared the learning effects among tracing, non-tracing, and cueing methods. In the cueing condition, they drew learners' attention to the same content as the tracing condition by highlighting the contents, where participants were asked to trace those contents in the tracing condition. Results showed that the tracing group had higher near transfer scores and far transfer scores than the cueing and no tracing group. The cueing group also performed better than the no tracing group on the far transfer tests, suggesting that cueing as an attention-managing method could facilitate learning. This experiment isolated the attention factor from the finger tracing, showing that finger tracing may benefit from attention enhancement near the hands, but it also has a complex mechanism other than attention only.

3.2.3 Tracing Facilitates Information Processing

As the previous two perspectives explained the tracing effect from the multi-sensory effect and hand proximity effect, the explanation from the cognitive load perspective focuses more on how tracing affects information processing, including information searching, packaging, and schema construction.

The first explanation under the cognitive load perspective focuses on information searching, which is associated with reducing extraneous cognitive load through finger tracing. Searching and matching relevant elements is a necessary process during learning, and the cognitive load imposed by this process may directly affect learning. Studies on the split-attention effect (Chandler & Sweller, 1992) demonstrated that presenting learning materials in an integrated format can reduce extraneous cognitive load. Similarly, finger tracing can support this necessary information searching and matching process and reduce the extraneous cognitive load by specifying learning materials areas and thus narrowing down the scope of searching and matching. This argument was supported by empirical results that tracing group reported significantly lower extraneous cognitive load (e.g., Ginns & King, 2021; Tang et al., 2019; Wang et al., 2022, 2025).

The second explanation under the cognitive load perspective focuses on information organising. The information-packaging hypothesis (Kita, 2000) holds that hand gestures can assist in organising information by packaging it into an single chunk (Alibali, 2005; Alibali et al., 2000), thereby decreasing the element interactivity and promoting information processing. F.-T. Hu et al. (2015) extended the above theorising to propose that tracing out specified textual and diagrammatic information would likewise act to chunk these elements into a unified schema, helping learners to construct higher-order schema (Ginns et al., 2020).

The third explanation under the cognitive load perspective focuses on information processing. The enhancement in attention near human hands was considered as a mechanism that could facilitate detailed assessment of objects in the visual field (Abrams et al., 2008), which is a biologically primary system that aims to prevent potentially survival-threatening danger, but also allows humans to evolve to more actively process the information near the hands. Thus, tracing as a biologically primary skill that requires hand movement may induce intensified information processing, resulting in enhanced schema construction and possibly automation.

3.3 Future Directions for Tracing Effect Research

Though the effect of tracing appears to be robust and various explanations are proposed to clarify its role in supporting learning, there is still much to learn about its underlying mechanisms and boundary conditions for effectiveness (Park et al., 2019). Researchers are continuously exploring the role of finger tracing across more fields, populations, and instructional designs. Beyond conventional exploration, researchers have also attempted to push the boundaries of tracing research from two perspectives: enhancing tracing effect via considering dosage, and combining tracing with other effects to further support learning.

3.3.1 Tracing's Dosage

Since tracing has been demonstrated to enhance learning, the question of whether more tracing yields better results has been an open question for researchers. A member of tracing studies have speculated that larger tracing actions may generate larger tracing effects (e.g., Ginns & King, 2021; Tang et al., 2019). By applying large-sized learning materials (A1-sized poster), Tang et al. (2019) demonstrated tracing's effect on learning materials of this size. Similarly, Ginns and King (2021) presented learning contents on a 23-inch computer screen and demonstrated tracing's effect. They speculated that larger media may trigger larger movements and hence promote larger effects on learning.

Considering the direct comparison between larger size finger tracing and smaller size finger tracing, Galbraith and Ginns (2023) conducted an experiment with adult learners. Participants were randomly allocated to an experimental (tracing out larger ellipses) or control (tracing out smaller ellipses) condition to learn a mental math strategy. However, participants in the control condition had better performance than those in the experimental condition after controlling for prior ability, especially for those participants with lower pre-test scores. Galbraith and Ginns speculated this pattern of results was due to a split attention effect, as tracing the larger ellipses may have guided participants' attention to areas away from key elements (numbers) in the worked examples. Zuo and Lin (2025) conducted experiments to compare finger tracing (smaller size) and whole-body tracing (bigger size). Participants in the whole-body tracing condition were presented with a learning poster placed on the floor and asked to trace the learning material with their feet by walking on the learning material. They hypothesised that the whole-body tracing, as a larger body movement, may lead to better effects. Though no difference was found in the learning performance between conditions, they found that whole-body tracing may enhance learners' interest (Experiment 2), and there were complicated interaction effects between condition and participants' perception of difficulty on three types of cognitive load.

Though the results of previous research have not fully supported hypothesis, they nevertheless demonstrated the potential for further exploration in this direction. The relationship between tracing dosage and learning outcomes remains a subject of ongoing research interest, which is explored as a supplementary analysis in this PhD study.

3.3.2 Compound Effects

Besides the dosage of tracing effect itself, the collaboration between tracing and other effects also constitutes a major focus of exploring the boundaries of tracing effects. Researchers have begun to focus on potential compound effects of tracing, or combined tracing with other cognitive load effects, to further explore its mechanisms.

As Sweller et al. (2019) argued, compound cognitive load effects and human movement research are worthwhile future directions for CLT. Conducting compound effects research is an important trend in CLT, as such research identifies other cognitive load effects' limits and thus promotes CLT's development as a theory. The element interactivity effect (Sweller, 1994), expertise reversal effect (Kalyuga et al., 2003), guidance-fading effect (Renkl & Atkinson, 2003), transient information effect (Leahy & Sweller, 2011), and self-management effect (Roodenrys et al., 2012) are all examples of compound effects discovered and developed in recent decades.

As for tracing, Wang et al. (2022) combined another cognitive load effect, the imagination effect (Cooper et al., 2001; Ginns et al., 2003), with finger tracing and found that sequencing tracing and imagination supported learning in both the acquisition phase and the post-test phase to a greater extent than an equivalent period spent tracing. Learner self-reported intrinsic motivation was also boosted in both tracing conditions. Zuo et al. (2025) considered the split-attention effect and finger tracing simultaneously, examining the effect of finger tracing on spatially separated or integrated learning materials. Though no significant difference was found between the two kinds of learning materials, their research expanded the field of tracing research and suggested the boundary conditions under which finger tracing could enhance learning. Previous research showed that tracing could be considered with other cognitive load effects to further explore its effect.

Understanding tracing's potential to support learning is far from being tapped; the current study speculates that the variability effect (Paas & Van Merriënboer, 1994) is similar to the imagination effect in being a cognitive load effect potentially acting as part of a compound effect with tracing. The combination of finger tracing and the variability effect is discussed in the next chapter.

Chapter Four: Tracing, Variability, and Current Research

Since sandpaper letters were introduced in Western educational settings by Montessori, exploration into the application of finger tracing to support learning has never ceased. Following extensive investigations into tracing effects in various disciplinary and age groups, recent research has begun to examine the combination of tracing and other instructional strategies. The current research follows in the footsteps of B. Wang et al. (2022) and Zuo et al. (2025), conducting experiments to explore the combination of tracing and variability. This chapter starts with a brief introduction on the characteristics of the variability effect (for a detailed introduction, see the subsection “The Variability Effect” in the previous chapter), focusing on its potential for combining with other instructional strategies. Subsequently, this chapter discusses the theoretical possibilities for combining tracing and variability, predicting how the combination of these two factors may affect learning. This chapter concludes by discussing how variable tracing might be operationalised, and clarifies the practical actions employed in the experiments in current research.

4.1 Characteristics of the Variability Effect

Since the first demonstration of the variability effect by Paas and Van Merriënboer (1994), CLT researchers have viewed variability as an effective factor in promoting germane load related to schema construction and automation processes (van Merriënboer & Sluijsmans, 2008). In contrast to most CLT-based research focusing on extraneous cognitive load, the variability effect works by manipulating intrinsic cognitive load and optimising it (Sweller et al., 2011). To enable the variability to fulfil its intended role in promoting deeper understanding, it is crucial to keep learners’ intrinsic cognitive load at an appropriate level, not too high that the total cognitive load exceeds a student's working memory limit, nor too low that variability-based instruction cannot achieve the desired effect. Researchers have identified two ways to prevent overloading when applying high variability instruction: controlling intrinsic cognitive load by choosing learners with an appropriate mastery level, and decreasing extraneous cognitive load by applying other instructional designs alongside variable instruction.

Researchers have tested the variability effect across learners with different prior abilities, finding that learners with higher prior abilities benefited more from high variability instruction. Likourezos et al. (2019, Experiment 2) found an interaction effect between levels of variability (high vs. low) and levels of learner expertise (novices vs. experts), suggesting that more experienced learners learned more from high rather than low variability tasks while less experienced learners learned more from low rather than high variability tasks. Learners with higher prior abilities perceived lower intrinsic cognitive load during the lesson, as they were less prone to feeling overwhelmed under high-variability instruction. Learners' ratings of difficulty also supported the argument that more experienced learners had sufficient available working memory capacity to process high-variability information.

Combining the variability effect and other cognitive effects can effectively increase intrinsic cognitive load whilst avoiding cognitive overload. Paas and Van Merriënboer (1994) combined variability and worked examples together, finding that high variability instruction only increases learning performance under the worked-example condition. The main effect of variability was not significant in the study, suggesting that high variability instruction alone is insufficient to promote learning. Likourezos et al. (2019, Experiment 1) conducted a similar experiment and tested the variability effect under two instructional guidance conditions (worked examples vs. unguided problem solving), finding a significant variability effect on transfer question performance regardless of levels of instructional guidance. When learners' total cognitive load does not exceed the working memory limit without incorporating any additional methods, it is possible to find a variability effect in both instructional guidance conditions.

The significant interaction between variability level and other variables (e.g., instructional type, expertise level) has supported the argument that high variability could further enhance learning when working memory capacity is sufficient to handle increased cognitive load. Finger tracing, as a verified learning strategy to reduce cognitive load and promote learning, is a reasonable candidate strategy to work in tandem with variability to further support learning.

4.2 Theoretical Justifications for Combining Tracing and Variability

As mentioned in the previous section, tracing supports biologically second knowledge learning through three possible paths, which are expanding working memory capacity (Ginns, 2005; Mousavi et al., 1995), guiding the learner's attention (Cosman & Vecera, 2010; Reed et al., 2006), and supporting cognitive processing (Hu et al., 2015). These mechanisms behind tracing share similarities or complementarities with those that promote learning through variability, suggesting potential avenues for the combination of these two factors.

Tracing's effective mechanisms are somewhat similar to variability's influential effect on learning. Firstly, they both can help learners notice the key learning point. The tracing intervention uses the index finger as an indicator to direct attention, utilising the characteristic that human vision tends to focus near the hands (Cosman & Vecera, 2010; Reed et al., 2006). The variability intervention shows variable examples with some constant elements, helping learners identify task-relevant dimensions and notice the core learning content (Marton, 2015; Marton & Booth, 1997; Raviv et al., 2022). Though implemented through different interventions, both tracing and variability aim to enable learners to focus on the learning content.

Secondly, tracing and variability both contribute to the information processing process. The tracing intervention supports searching and matching processes, while variability promotes retrieval and reconstruction processes. Finger tracing specifies the learning areas for learners, narrowing down the scope of searching and matching. Variability provides more types of learning examples, enhancing the retrieval process through a more effortful cycle of forgetting and reconstruction. These potential similarities in mechanisms indicate that the combination of tracing and variability is possible and significant.

The possibility that tracing and variability can compensate for each other also increases their combination potential. Paas and Van Merriënboer (1994) concluded that the positive effects of variability on schema acquisition would manifest only when extraneous cognitive load is reduced. Correspondingly, tracing can reduce extraneous cognitive load by narrowing down the searching and matching areas. When the extraneous cognitive load is reduced, learners' working memory capacity is available for intrinsic load, allowing learners to engage in more complex learning and deeper processing. Furthermore, tracing can expand extra working memory capacity by adding the haptic channel of working memory. In summary, tracing may compensate for variability's potential risk of overwhelming cognitive load, indicating that combining them may further promote learning than single tracing or single variability intervention.

4.3 Operationalising Variable Tracing Actions

Theoretical mechanisms behind tracing and variability open the possibility of combining these two factors; practical way to combine tracing and variability is discussed in this subsection.

Among the four types of variability (for a review, see Raviv et al., 2022), most previous research from the CLT perspective focuses on the content difference, that is, the numerosity difference and heterogeneity among learning materials (e.g., Paas & Van Merriënboer, 1994, variability of worked examples). In contrast to incorporating variability across multiple worked examples, tracing focuses on external motor activity when engaging with lesson materials (such as worked examples) rather than the learning material itself, suggesting that the specific type of variability applicable to tracing research is closer to the situational (contextual) difference and scheduling arrangement (e.g., spacing, interleaving) forms of variability.

CLT research on interleaving (Chen et al., 2021) proposed the discriminative-contrast hypothesis, that interleaving instructional design enhances learning by assisting learners in discriminating between learning topics. As to the present research on tracing, the interleaving intervention and the discriminative-contrast hypothesis can be expanded to alternating two index fingers to tracing, which may help learners to discriminate different solution steps within a worked example. At the moment that learners are asked to change the hand they use to trace, they are simultaneously being reminded that this is a crucial point of the learning material. A learner's attention is directed towards observing the differences between the learning content before and after the finger switch. This alternating tracing approach can also pull learners out of the habitual motion of single finger tracing, preventing single finger tracing from becoming an unconscious and meaningless action. Therefore, the present research aims to investigate variability in tracing by focusing on alternating which index finger traces a specified step in a worked example.

Chapter Five: Methodology

5.1 Aim and Hypotheses

The current research was based on previous findings that finger tracing can improve learning across various subject areas, media, and age groups. It aimed to find how tracing impacts learning when variability is combined with tracing, comparing the classic non-alternating finger tracing strategy and the alternating fingers tracing strategy's effect on promoting learning. The entire research was approved by and conducted under the supervision of The University of Sydney Human Research Ethics Committee (HREC), with project no. 2023/397.

Three experiments were conducted to test the impacts on learning outcomes as well as on intrinsic motivation and cognitive load. Experiment 1 investigated whether alternating tracing could further support learning. Experiment 2 revised part of the learning materials to reduce a potential split attention effect and extraneous cognitive load found in Experiment 1. Experiment 3 built on the previous studies, but divided the learning process into two stages to obtain comparative data from the two stages, reflecting more detailed results on the difference between the non-alternating finger tracing strategy and the alternating fingers tracing strategy.

In each experiment, there were two conditions, including a control group and an experimental group. In Experiment 1 and Experiment 2, participants in the control group (non-alternating group) applied the non-alternating finger tracing strategy, using their dominant hand's index finger to trace learning materials. Participants in the experimental group (alternating group) applied the alternating fingers tracing strategy, tracing the learning materials alternating between their index fingers of both hands. In Experiment 3, participants in both groups applied the same non-alternating finger tracing strategy across the first stage and applied different tracing strategies according to their conditions at the second stage. Each experiment included two days of participation. The first day's experiment began with the consent phase, followed by the demographics phase, the pre-test phase, the learning phase, the self-reported phase, and the immediate post-test phase. The second day's experiment contained the delayed post-test phase and the debriefing phase.

The hypotheses below were examined via Experiment 1 and Experiment 2, presented across the next two chapters:

H1: Participants' problem-solving performance in the acquisition phase will differ across conditions, reflected by problem-solving accuracy and problem-solving time. Specifically, participants in the non-alternating group will solve fewer problems and/or solve problems more slowly than participants in the alternating group.

H2: Participants' self-reported intrinsic motivation will differ across conditions. Specifically, participants in the non-alternating group will report lower intrinsic motivation than participants in the alternating group.

H3: Participants' self-reported intrinsic cognitive load will differ across conditions. Specifically, participants in the non-alternating group will report lower intrinsic cognitive load than participants in the alternating group.

H4: Participants' self-reported extraneous cognitive load will not differ across conditions.

H5: Participants' self-reported germane processing load will differ across conditions. Specifically, participants in the non-alternating group will report a lower germane processing load than participants in the alternating group.

H6: Participants' problem-solving performance in the immediate post-test phase will differ across conditions, reflected by problem-solving accuracy and problem-solving time for total test problems. Specifically, participants in the non-alternating group will solve fewer problems and/or solve problems more slowly than participants in the alternating group.

H7: Participants' problem-solving performance in the delayed post-test phase will differ across conditions, reflected by problem-solving accuracy and problem-solving time for total test problems. Specifically, participants in the non-alternating group will solve fewer problems and/or solve problems more slowly than participants in the alternating group.

The hypotheses of Experiment 3 were a little different from Experiment 1 and Experiment 2 because of its adjustments to experimental procedures. The specific hypotheses of Experiment 3 were presented at the beginning of the experiment chapter.

5.2 Experimental Design

All three experiments in the current research followed the same design. In addition to typical design principles used in CLT experiments, several factors or variables were added to the current research to obtain richer and more accurate experimental data.

5.2.1 Test Accuracy and Test Time to Solution

Two indicators were used to evaluate participants' learning performance in the current research. Besides the test score, the test time to solution was also applied. Though test accuracy can directly reflect participants' mastery of learning content, test time to solution can further differentiate participants' mastery level when they have equal test accuracy. Test time to solution was also used to evaluate participants' learning performance when a ceiling effect was found on test scores. This design referred to Wang's (2022) research, which was also a finger tracing research in the CLT field. Besides its differentiating function, the test question time to solution also had a more detailed and discriminative distribution than the test score, which was more suitable for statistical analysis.

However, when there is a significant difference in test accuracy between participants, the test time solution becomes meaningless. Comparing the mastery level of a participant who answers correctly at a slower pace with a participant who answers incorrectly at a faster pace, there is no doubt that the participant who answers correctly has a better learning performance. Thus, participants' test accuracy was the first priority indicator to reflect participants' learning performance in the current research; the test time to solution was an alternative measure when no difference between conditions was found in participants' test accuracy.

5.2.2 Delayed Post-Test Design

In addition to including acquisition stage performance and post-study performance as in typical CLT experiments, the current research also included a delayed post-test as another variable to evaluate participants' performance.

A delayed post-test is considered to more accurately reflect true mastery of the learning content than the immediate test, because it gives participants the opportunity to forget some information and avoid inferring answers (Andre, 1990). Immediate post-test results "may be influenced by many transient factors such as fatigue, boredom or excitement" (Dubrowski et al., 2010, p. 102). Some learning methods have been found to have a more substantial effect on the delayed test compared to the immediate test (Andre & Thieman, 1988; Rickards, 1976). Focusing on CLT research, while typical studies use only an immediate post-test, a delayed test has been applied in some studies, such as investigations of the split attention effect (e.g., Chandler & Sweller, 1992, Experiment 1) and the generation effect (Chen et al., 2016).

Hand gesture, as a typical sensorimotor experience, has also been found to significantly promote learning on a delayed test but not an immediate test (Cook et al., 2008). A delayed post-test design can also be found in tracing research. Yeo (2024) applied a four-week delayed post-test to compare the tracing and the non-tracing conditions' learning supportive effect on learning worked-example mathematical problems. Considering variability-related research, the effect of interleaved practice was also only found on the delayed test in several studies (e.g., Mielicki & Wiley, 2022; Paulo & Robert, 2014; Taylor & Rohrer, 2010). In summary, delayed post-tests are desirable in educational research because they support more stringent tests of hypotheses, and were included in each of the current study's experiments.

Chapter Six: Experiment 1

6.1 Method

6.1.1 Sample Size and Participants

Power Analysis. According to statistical principles and long-standing research experience, numerous educational researchers suggest that at least 15 participants should be included in each condition in experimental research (Brysbaert, 2019; Cohen et al., 2007; Gall et al., 1996). Furthermore, the Central Limit Theorem holds that the sampling distribution can be considered normal when the sample size is larger than 30 (Kwak & Kim, 2017), which is important for later statistical analysis. To further clarify the effective sample size for the current research, we conducted a power analysis. The power analysis referred to the variability effect size (d = 0.57) in lab studies estimated from the meta-analytic findings of the contextual interference literature (Brady, 2004). One-tailed tests were applied because the present research fitted the second condition of Kimmel's (1957) criteria for using one-tailed tests. The current research aimed to test the effect of alternating finger tracing compared to the "baseline" (non-alternating) instructional design tested in previous studies. Therefore, directional hypotheses were more efficient and appropriate in this study. When using directional hypotheses, the power for the one-sided test can be calculated with a doubled alpha level parameter (Caldwell et al., 2022). Thus, the power analysis was carried out as a one-tailed test, with effect size d = 0.57, α = 0.05 and desired power = 0.90. The result showed that the current research required at least 54 participants for each group, which is 108 participants in total.

Participants. The first experiment was conducted from August 2023 to October 2023 at East

China Normal University (ECNU), Shanghai, China. All participants are ECNU or nearby college students and teachers. A total of 115 participants attended the experiment, and all of them completed the whole experiment. All participants were educated to at least the undergraduate level (including those who were currently in undergraduate programs) and were able to read English in paragraphs. All participants were randomly assigned to the control condition (applying a non-alternating finger tracing strategy, N = 58) and the experimental condition (application of an alternating fingers tracing strategy, N = 57). The gender ratio of the participants was not balanced (90 females, 24 males, 1 preferred not to say), but it was consistent with the distribution of the population where the data were collected. Participants' ages ranged from 18 to 29 years old (M = 22.52, SD = 2.52).

6.1.2 Materials and Procedure

The recruited participants participated in the experiments individually in a quiet room with the guidance from the PhD researcher (Xufei Zhang). The experiment contained a delayed post-test, which required participants to come back again on the next day to complete it. Paper-based materials used in the present study were chosen for their suitability for a laboratory-based learning experiment. Such experiments require lesson and test materials focused on a topic that is novel for learners, but is learnable within a short period of time, as well as being sensitive to instructional design manipulations. The mental mathematics lesson materials used in the present series of experiments were adapted from previous studies investigating a range of instructional redesigns (e.g., Ginns et al., 2019; Smyrnis & Ginns, 2016; Wang et al., 2022). The full set of experimental materials is given in Appendix A. The whole experiment needed about 30-40 minutes, with the first day lasting 25-30 minutes and the second day lasting 5-10 minutes. The first day's experiment started with the consent phase, followed by the demographics phase, the pre-test phase, the learning phase, the self-reported phase, and the immediate post-test phase. The second day's experiment contained the delayed post-test phase and the debriefing phase.

Participants in both conditions received identical instructions except for the learning phase, in which they were told to use different tracing strategies (according to their condition). To avoid ambiguity caused by translation, all the materials mentioned above were paper-based and presented in printed English (adapted from materials used in Wang et al., 2022). Though English was not the first language of the participants, English as a compulsory course in the national college entrance exam had been studied by all participants for at least 10 years (from Grade 3 to Grade 12). With years of compulsory English courses, all participants were able to read English in paragraphs. Additional oral instruction between phases was given in the participants' first language (Mandarin). Participants were also encouraged to ask in Mandarin if they had any doubts about the experiment. The whole experiment was videotaped after gaining participants' consent. Recorded participants' tracing actions were counted as a check on treatment fidelity (Mowbray et al., 2003), in line with previous CLT research (Eielts et al., 2020; Galbraith & Ginns, 2023).

Consent Phase. Participants were notified of basic information about the experiment and signed a consent form. To avoid the subject-expectancy effect (Rosenthal & Fode, 1963), the current research hid the real purpose of the experiment from participants. Participants were not informed that the experiment was designed to research the tracing strategy. Instead, they were told that this experiment was used to research a new method of two-digit multiplication.

To randomly assign participants to experimental conditions, participants with odd-numbered positions in the experiment sequence were asked to flip a coin to determine their condition assignment. If the coin landed heads up, the participant was assigned to the non-alternating group, and if tails, the participant was assigned to the alternating group. Participants with even-numbered positions in the experiment sequence did not need to flip a coin, and they were automatically assigned to the opposite condition of the preceding participant.

The video camera was aimed at the desktop portion and only captured participants' hand movements. Video recording was commenced after the participant signed the consent form.

Demographics Phase. The experiment officially started with participants answering a survey about their basic demographic information. Their gender and age were collected by choice and fill-in-the-blank questions. A handedness survey (Nicholls et al., 2013) with ten questions followed. Participants reported their hand preference in ten different situations (e.g., In which hand do you prefer to hold a toothbrush when cleaning your teeth? When buttering bread, which hand holds the knife?) to reflect their propensity to use one or both hands.

Pre-Test Phase. The pre-test phase was designed to assess participants' prior knowledge, in the form of arithmetical problem-solving skills. A calculation fluency test (Sowinski et al., 2014) was applied in this phase. The test consisted of three pages, and each page had 60 calculation questions on addition, subtraction, and multiplication separately.

Participants were given 1 minute for each page, and they were asked to answer as many questions as possible. Before participants started answering them, they were also told that the number of questions far exceeded the number that could be completed in 1 minute, so they did not need to worry about not being able to finish them. The number of correct questions answered by the participant on each page was recorded separately, and this result was treated as the participant's prior calculation ability.

Learning Phase. The learning material showed participants how to quickly calculate a two-digit multiplication, based on a method in Julius (1992, pp. 83–84), and consisted of 10 pages of instruction, including two pages of introduction and eight pages of instruction on the two-digit multiplication method.

The two-page introduction material told participants the procedure of this phase and the requirements of finger tracing while learning, and asked them to practice tracing with their fingers before instruction. The first page was the introduction, and the second page was the tracing practice. The introduction materials were identical for participants in both conditions, except that they were told that they needed to use their non-alternating finger or alternating fingers (according to the conditions) to trace the learning material.

The eight pages of instruction included four worked examples interleaved with four practice questions. All the worked examples were about the same kind of method, but the worked examples got progressively more difficult. The worked examples were adapted from Wang et al. (2022), and the current research adjusted some specific numbers in the worked examples to keep them at the appropriate difficulty level. The first three worked examples contained one carry-on calculation during the solving process (see Figure 1), and the involved numbers became progressively larger. The fourth worked example contained two carry-on calculations in the process (see Figure 2), which made it more difficult. For each worked example, the two-digit multiplication problem was solved in three steps, and instruction was provided at each step. The instruction included a short sentence explaining the step's calculating process, colour-coded horizontal and vertical formulas with key numbers in them, and colour-coded ellipses displaying the key numbers in the vertical formula. The correct answer to the whole worked example was shown at the end of the three steps. Beside each step, there was a separate text box that asked participants to trace the key ellipses with the requested index finger. The learning materials were identical for both conditions except for the separate text box, in which participants were asked to use their writing hand's index finger or their non-writing hand's index finger (according to the conditions) to trace the learning material. Participants in the non-alternating tracing condition were asked to use their writing hand's index finger to trace the learning material all the time, while participants in the alternating tracing condition were asked to switch their writing hand and non-writing hand's index finger to trace the given step.

Figure 1

A Worked Example Containing One Carry-on Calculation in Experiment 1

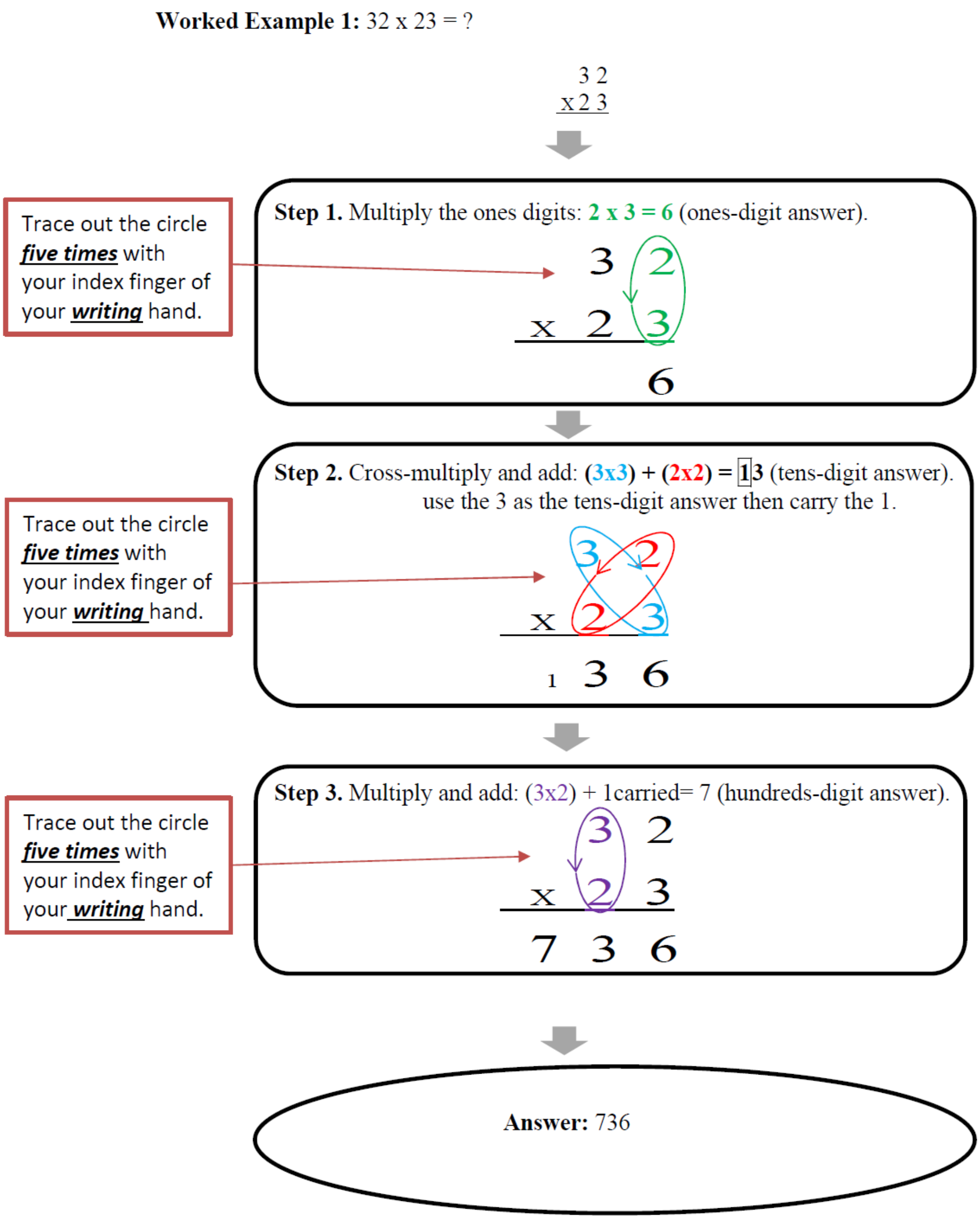
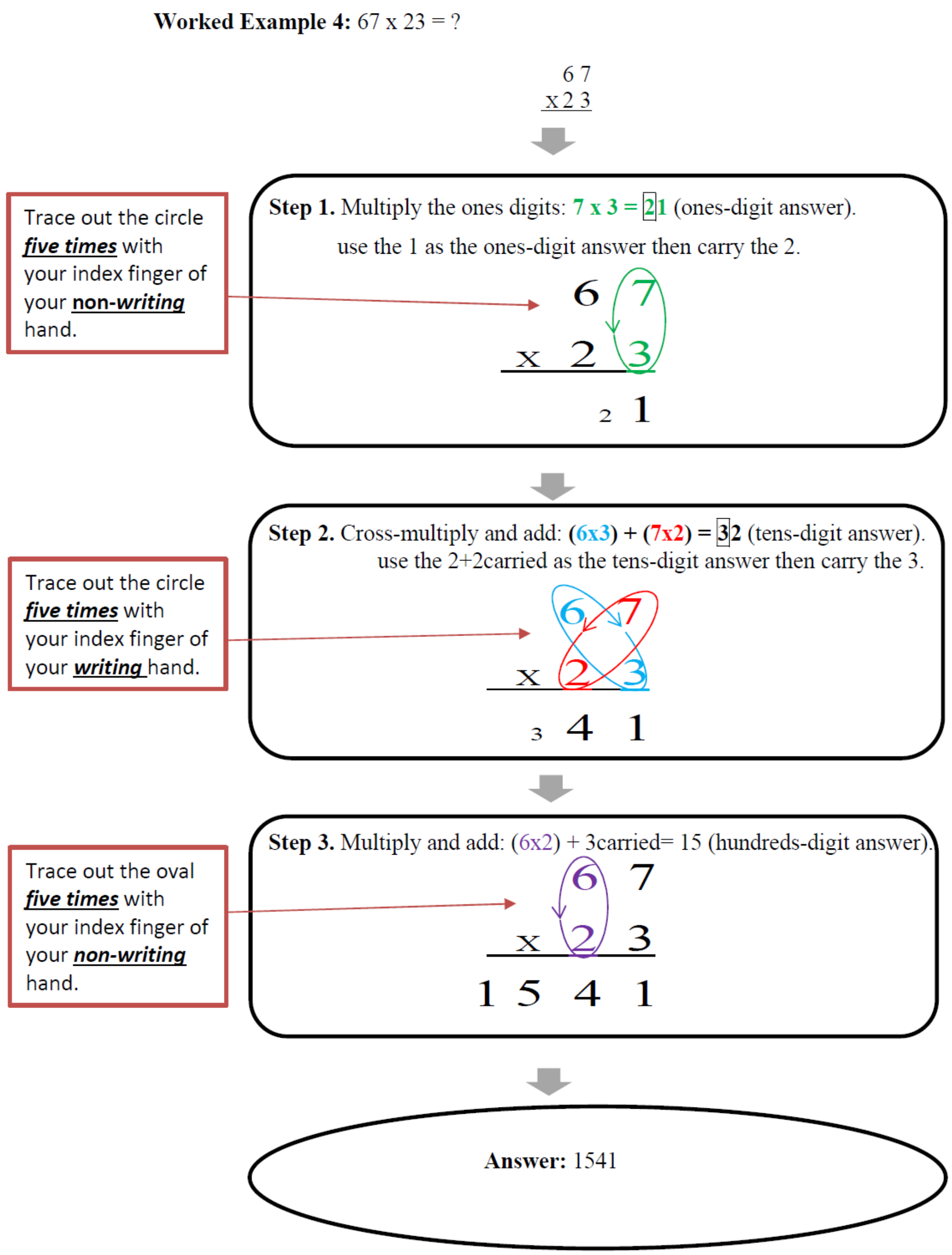


Figure 2

A Worked Example Containing Two Carry-on Calculations in Experiment 1



At the beginning of this phase, participants received a two-page instruction telling them how to learn and trace the learning materials in this phase. The researcher also gave oral instructions to help participants understand the procedure and the requirements in this phase. Participants could take their time to read the instructions until they fully understood the experiment's requirements, and they were encouraged to ask questions for clarification if needed. They were then asked to trace the ellipses on the tracing practice page with guidance from the researcher. Participants then started to learn the two-digit number multiplication method by studying the four worked examples one by one and answering the practice questions that followed each worked example. Participants had two minutes to study each worked example and 30 seconds to answer the following practice question of the corresponding difficulty. The researcher told participants whether their answer was correct or incorrect immediately after participants wrote down the answer. If the answer was wrong, then the participant continued to try until 30 seconds ran out. The sequence was to learn one worked example for two minutes, then answer one question, and learn the next worked example for another two minutes, then answer another question. During the two-minute study period, participants in both conditions traced the ellipses in the worked example steps with the specified index finger as the separate text box indicated. The tracing instructions in the two conditions were different, specifying the non-alternating index finger (writing hand's index finger) for the non-alternating tracing condition and using alternating index fingers (switching between writing and non-writing hands' index fingers) for the alternating tracing condition. During participants' learning time, the researcher sat to one side and watched participants' learning behaviour, reminding them when necessary to focus on learning and use the correct index finger to trace.

Self-report Phase. After the learning phase, participants completed a twenty-item survey on their motivation and different types of cognitive load. Participants needed to rate each statement from 0 (not at all applicable) to 8 (fully applicable), according to their experience in the learning phase. The intrinsic motivation scale was developed by McAuley et al. (1989), including five different items (e.g., "I enjoyed doing this activity very much"). The cognitive load scale (15 items) was developed by Krieglstein et al. (2023), including five items on intrinsic cognitive load (e.g., "The learning content was difficult to understand"), five items on extraneous cognitive load (e.g., " The design of the learning material made it difficult to find relevant information quickly"), and five items on germane processing (e.g., " I actively reflected upon the learning content"). When some participants asked the precise meaning or the referred-to item of some words and phrases in the statement, the researcher read a pre-prepared standardised translation of the statement in response to such questions from the participants.

Immediate Post-Test Phase. Participants needed to answer 20 two-digit multiplication questions in this phase. For each question, they had 20 seconds to answer, writing down the final answer within the time limit. The researcher did not tell the participants if their answers were correct or not. Participants who finished answering one question just moved to the next one, and the researcher started rerecording the time of the next question. The accuracy of participants' answers and the time they spent solving the questions were recorded to reflect their learning performance. Writing an answer that exceeded 20 seconds was recorded as "out of time", and even if the answer was correct, this question was seen as an unsuccessful answer, which was equal to wrong. The questions in the post-test phase involved larger numbers, which made the post-test questions a little more difficult than the practice questions in the learning phase.

Delayed Post-Test Phase. The delayed post-test phase was applied one day (approximately 24 hours) after the initial experiment. Participants needed to answer 20 two-digit multiplication questions again in the delayed post-test phase. The questions here were exactly the same ones as presented in the immediate post-test.

Debriefing phase. The researcher debriefed the whole experiment to the participants in this phase, clarifying the true purpose of the experiment. Participants were encouraged to express their feelings and thoughts about the experiment and its theoretical assumptions. Some participants asked about their accuracy in the test, and some were interested in the research design. The researcher responded to the participants' questions as much as possible without posing privacy concerns.

6.1.3 Data Processing and Model Check

One hundred and fifteen participants joined and completed the experiment, and these data constituted the full dataset. As participants' behaviours were fully videotaped, the video recordings can be used to check participant compliance with instructions. The fully compliant participants constituted the compliant dataset.

To check participants' compliance, a coding scheme was developed to quantify participants' effective tracing activities (see Appendix B). Their tracing actions on the required learning materials were recorded as valid counts. Some participants were excluded from the compliant dataset for the following two reasons.

The first reason was that participants' valid tracing actions were insufficient. Each worked example included four ellipses in all three steps. The learning material asked participants to trace five times for each ellipse. Thus, if participants followed the instruction strictly, then they should trace at least 20 times for each worked example and 80 times in total. However, in the actual learning process, participants might be confused about their tracing times and unable to trace a complete ellipse in the final ellipse. Considering this, we set 64 times in total (four times for each ellipse) as a pass line. Any participant traced less than 64 times in total was deemed "non-compliant" and excluded. Any participant who traced insufficiently (less than 16 times for each worked example) in two or more worked examples was also excluded.

The second potential reason for exclusion was when a participant's tracing action did not fulfil the group condition. Though the instructions on the learning material described the requested tracing action in detail, some participants did not read carefully and did not follow the instructions. Thus, some participants traced the learning material with alternating fingers, even if they were assigned to the non-alternating tracing group, or vice versa; those participants were excluded. Some participants in the alternating tracing group may have been confused about which hand to trace and used the wrong one at the beginning, but they adjusted their actions later, and their tracing actions were generally compliant with alternating tracing requirements. So, these participants were retained in the following analysis. Only those participants who did not meet the group condition are excluded.

Ninety-five participants were retained after the compliance check, and 20 participants were excluded. Fifteen participants were excluded due to insufficient tracing times (participant number 8, 15, 16, 17, 18, 23, 24, 29, 35, 54, 75, 80, 88, 93, 112), and four participants were excluded because of incorrect tracing action (participant number 43, 48, 68, 105); another participant was excluded because of data corruption (participant number 46, 30 seconds video recording was lost in Worked Example 4).

The following data analysis was based on the compliant dataset (95 participants' results). The majority of statistical analyses were conducted via jamovi (2.6.26) (The jamovi project, 2024). Detailed data processing procedures were separately presented for each variable.

Handedness Score (Hand Preference and Coordination). Participants' hand preference was collected via the handedness survey (Nicholls et al., 2013). As Figure 3 showed, all participants needed to report their propensity to use one or both hands. A simple coding was applied to the survey and participants' answers. The answer "left" was coded as "-1", either was coded as "0", and "right" was coded as "1". A handedness score between -1 to 1 was yielded for each participant by averaging the coding of ten questions. Scores closer to 0 indicate greater balance in participants' hand use. Conversely, scores closer to -1 indicate stronger left-handedness in participants, while those approaching 1 reflect right-handed preference. The absolute value of the handedness score was also calculated to reflect participants' hand coordination.

Figure 3

The Handedness Survey

Please choose your preferred hand in the following situations

Please tick one box for each question, indicating whether you prefer to use the left-hand, either-hand, or the right-hand for that task. Only tick the 'either' box if one hand is truly no better than the other. Please answer all questions, and even if you have had little experience in a particular task, try imagining doing that.

		Left	Either	Right
1	With which hand do you write?			
2	In which hand do you prefer to use a spoon when eating?			
3	In which hand do you prefer to hold a toothbrush when cleaning your teeth?			
4	In which hand do you hold a match when you strike it?			
5	In which hand do you prefer to hold the rubber when erasing a pencil mark?			
6	In which hand do you hold the needle when you are sewing?			
7	When buttering bread, which hand holds the knife?			
8	In which hand do you hold a hammer?			
9	In which hand do you hold the peeler when peeling an apple?			
10	Which hand do you use to draw?			

A reliability analysis was conducted (Revelle, 2024) to check the consistency of participants' responses to the handedness survey. As the first item in the handedness survey asked participants' hand preference for writing, and all of them chose "the right hand", there was no variance on this item. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that the remaining nine items from the handedness survey had suitable levels of reliability (Cronbach's $\alpha = .82$; McDonald's $\omega = .86$).

Pre-Test Score (Prior Calculation Ability). Participants' prior abilities were evaluated via their performance in the calculation fluency test. Participants finished a three-page calculation test on addition, subtraction, and multiplication (one test per page). Every correctly answered question on these pages was counted as one point, producing three separate scores for addition, subtraction, and multiplication fluency.

Table 1 presents participants' performance in the calculation fluency test. On average, participants gained a higher score on addition than subtraction, and both of these scores were higher than multiplication. These results were consistent with Sowinski et al.'s (2014) results. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that the pre-test scores had suitable levels of reliability (Cronbach's $\alpha = .86$; McDonald's $\omega = .94$).

Though the following learning task in this experiment only involved addition and multiplication in the calculation process, including subtraction scores together could make the total scores more widely distributed, and reflect participants' calculation ability at a more general level. Scoring the test as the total number of correctly solved items across the three pages was also suggested by the test developer (Sowinski et al., 2014). Thus, three separate scores were added together to generate an overall score to represent their general calculation ability in the following analysis.

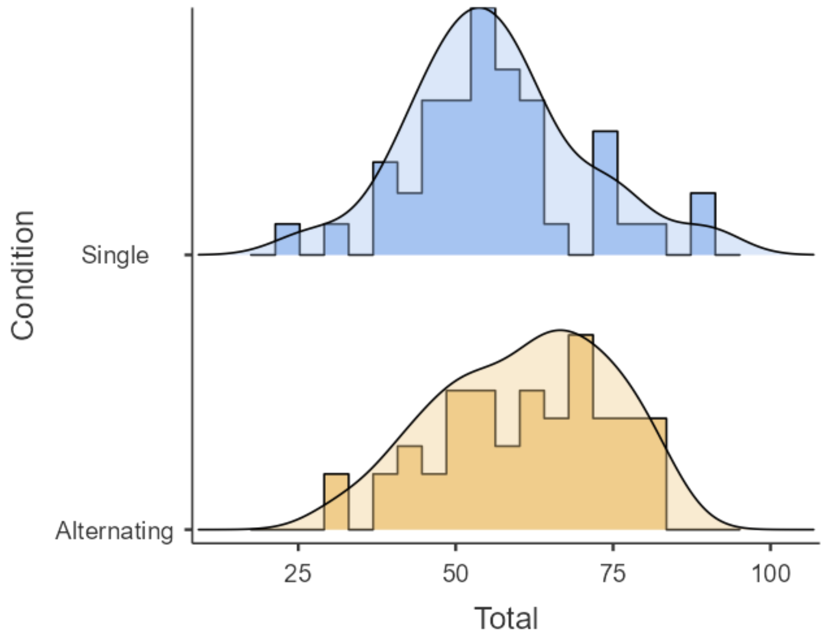
Table 1

Participants' Performance in the Calculation Fluency Test in Experiment 1

Test Type	n	*M*	*SD*
Addition (/60)	95	23.5	5.7
Subtraction (/60)	95	19.0	5.8
Multiplication (/60)	95	16.3	4.0
Total (/180)	95	58.8	13.8

Figure 4

The Distribution of Participants' Total Scores in the Pretest in Experiment 1



In this experiment, participants' prior abilities met the assumption of being normally distributed, based on the Shapiro-Wilk (1965) test. This test was considered the most appropriate method for small sample sizes (<50) and was equally effective compared to other methods for medium-sized samples ($50 \leq n < 300$) (Mishra et al., 2019). As this experiment had 95 participants in total and 45 or 50 participants in each group, the Shapiro-Wilk test was considered a suitable choice for the current study when the normality analyses involving subgroup data. The overall distribution (Shapiro-Wilk $W = .99$, $p = .861$) as well as the distribution of each condition satisfied the normal distribution. As Figure 4 shows, the non-alternating tracing group (Shapiro-Wilk $W = .98$, $p = .458$) was more normally distributed than the alternating tracing group (Shapiro-Wilk $W = .97$, $p = .200$).

Practice Phase Learning Performance. Participants' learning performance in the learning phase was evaluated via two indicators: the number of practice questions they answered correctly and the time they used to answer the questions. The former was defined as the practice score, and the latter was defined as the practice test time to solution. The practice score was an integer that ranged from 0 to 4. The practice time to solution was obtained by adding up the time participants used to correctly answer each practice question. If the final answer was wrong or participants couldn't give an answer within the time limit (30 seconds), the time of 30 seconds was recorded. Practice time to solution therefore ranged from 0 to 120 seconds.

Post-Lesson Self-Report. Participants' self-reports reflected four variables: intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing, each evaluated by five separate statements. The score ranged from 0 to 8, reflecting participants' agreement with the statement.

To check the construct validity of the current survey results, a confirmatory factor analysis (CFA) was conducted using jamovi (Rosseel et al., 2024). Though CFA's result was generally considered to be reliable only with a large sample size ($n > 300$) (Comrey & Lee, 1992), Wolf et al. (2013) proposed the possibility that a CFA model with high factor loading might be applicable to a minimal sample size lower than 100. Thus, we continued to conduct the CFA model with the current dataset. To check the assumption of multivariate normality, the squared Mahalanobis distances were calculated for the data and plotted against the quantiles of a Chi-square distribution (DeCarlo, 1997). As Figure 5 indicates, the assumption of multivariate normality was largely met. Multicollinearity was assessed via the squared multiple correlations and the determinant of the correlation matrix, not finding any variables had an R^2 > .90, which indicated low risk in multicollinearity (Kline, 2015).

Indices of model fit including the Comparative Fit Index (Bentler, 1990), Tucker-Lewis Index (Tucker & Lewis, 1973) and Root Mean Square Error of Approximation (Hu & Bentler, 1999) all showed the model was a mediocre fit (CFI = .87, TLI = .85, RMSEA = .100). Further checking the factor loadings (see Table 2), the indicators "GCL_2" and "ICL_5" were found to be unsuitable for the whole model. They were excluded because they were not significantly related to the main factors. The model was rerun without these two indicators, and the results showed a better model fit (CFI = .89, TLI = .87, *RMSEA* = .106). These two indicators were excluded from all the following analyses.

Figure 5

Mahalanobis Distance Scatterplot Testing Multivariate Normality in Experiment 1

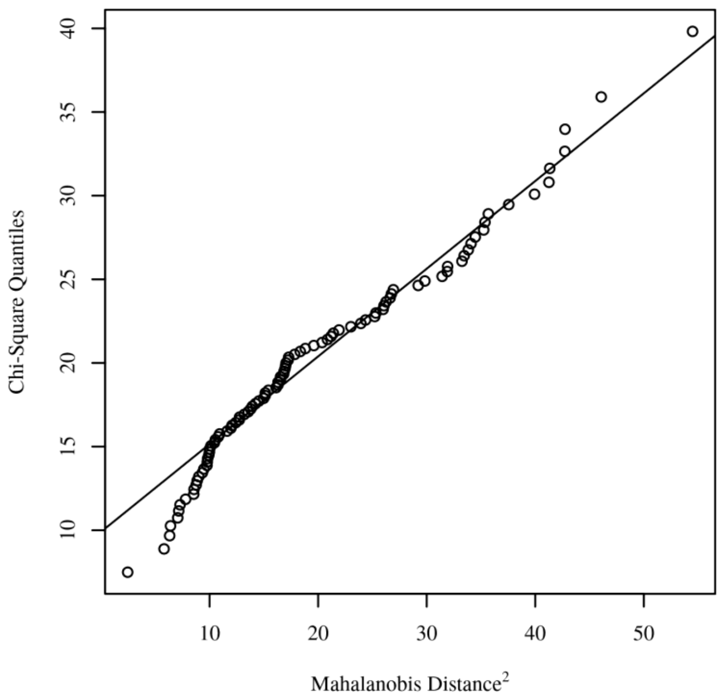


Table 2

Original Factor Loadings in Experiment 1

Factor	Indicator	Estimate	SE	Z	*p*	Stand. Estimate
Intrinsic Motivation	Motivation1	1.22	0.12	10.40	< .001	0.86
	Motivation2	1.60	0.13	12.38	< .001	0.95
	Motivation3	1.69	0.14	11.97	< .001	0.93
	Motivation4	1.63	0.13	12.39	< .001	0.95
	Motivation5	1.24	0.20	6.06	< .001	0.58
Intrinsic Cognitive Load	ICL_1	1.09	0.10	11.22	< .001	0.90
	ICL_2	0.99	0.08	12.45	< .001	0.96
	ICL_3	0.87	0.12	7.21	< .001	0.67
	ICL_4	0.87	0.12	7.09	< .001	0.67
	ICL_5	0.39	0.25	1.55	.122	0.16
Extraneous Cognitive Load	ECL_1	0.85	0.11	7.60	< .001	0.71
	ECL_2	1.26	0.16	7.85	< .001	0.73
	ECL_3	1.42	0.14	9.83	< .001	0.85
	ECL_4	1.47	0.16	9.24	< .001	0.81
	ECL_5	1.39	0.19	7.50	< .001	0.70
Germane Processing	GCL_1	0.88	0.17	5.35	< .001	0.59
	GCL_2	0.28	0.23	1.23	.218	0.15
	GCL_3	0.95	0.15	6.43	< .001	0.66
	GCL_4	0.96	0.15	6.54	< .001	0.68
	GCL_5	1.08	0.17	6.29	< .001	0.67

Reliability tests were conducted (Revelle, 2024) to check participants' answers' consistency and dependability. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that four separate factors had suitable levels of reliability (see Table 3).

Participants' motivation and cognitive load were evaluated via the aforementioned four factors and 18 items. The average score of each factor was used for the following analyses. All scores ranged from 0 to 8, and higher scores indicate higher motivation or cognitive load.

The Shapiro-Wilk (1965) test was applied to check the normality of participants' average intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing score.

Table 3

Reliability Scores in Experiment 1

Variable	Cronbach's α	McDonald's ω
Intrinsic Motivation (5 items)	.92	.93
Intrinsic Cognitive Load (4 items)	.88	.89
Extraneous Cognitive Load (5 items)	.86	.87
Germane Processing (4 items)	.74	.75

Participants' average intrinsic motivation (Shapiro-Wilk $W = .96$, $p = .005$) was not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic motivation distribution show a pronounced left skew: the 25^{th} percentile is 4.6 while the median is as high as 5.8. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic motivation, participant responses of 8, the highest possible response, constituted 18.9%, 24.2%, 21.1%, 15.8%, and 10.5%, for items 1-5 respectively. Based on McHorney and Tarlov's (1995) guidance that ceiling or floor effects are likely to be an issue when more than 15% of respondents respond with either the highest or lowest possible score, a ceiling effect was evident for intrinsic motivation.

Participants' average intrinsic cognitive load (Shapiro-Wilk $W = .88$, $p < .001$) also were not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic cognitive load distribution show a pronounced right skew: the median is 0.8 while the 75^{th} percentile is 1.8. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic cognitive load, participant responses of 0, the lowest possible response, constituted 36.8%, 40.0%, 45.3%, and 42.1%, for items 1-4 respectively; thus, a floor effect was evident for self-reports of intrinsic cognitive load (McHorney & Tarlov, 1995).

Participants' average extraneous cognitive load (Shapiro-Wilk $W = .91$, $p < .001$) also were not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the extraneous cognitive load distribution show a pronounced right skew: the median is 1.4 while the 75^{th} percentile is 2.5. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For extraneous cognitive load, participant responses of 0, the lowest possible response, constituted 21.1%, 23.2%, 29.5%, 32.6% and 32.6%, for items 1-5 respectively; thus, a floor effect was evident for self-reports of extraneous cognitive load (McHorney & Tarlov, 1995).

Participants' average germane processing score (Shapiro-Wilk $W = .95$, $p = .002$) also was not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the motivation distribution show a pronounced left skew: the 25th percentile is 5.5 while the median is as high as 6.5. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For germane processing, participant responses of 8, the highest possible response, constituted 12.6%, 21.1%, 27.4%, and 26.3%, for items 1 and 3-5 respectively; thus, a ceiling effect was evident for self-reports of germane processing (McHorney & Tarlov, 1995).

Immediate Post-test Phase Learning Performance. Participants' learning performance in the immediate post-test phase was evaluated via two indicators: the number of test questions they answered correctly, and the time they used to answer the questions. The former was defined as the immediate post-test score, and the latter was defined as the immediate post-test time. The immediate post-test score was an integer that ranged from 0 to 20. The immediate post-test time was obtained by adding the time participants used to correctly answer each test question. If a participant could not give a correct answer within the time limit (20 seconds), then the maximum time of 20 seconds was recorded. The immediate post-test time to solution thus ranged from 0 to 400 seconds.

Delayed Post-test Phase Learning Performance. Participants' learning performance in the delayed post-test phase was evaluated via the same indicators as used in the Immediate Post-test. Thus, the delayed post-test score was an integer that ranged from 0 to 20. As for the Immediate Post-test to solution, the Delayed Post-test time to solution was obtained by adding the time participants used to correctly answer each test question. If a participant could not give a correct answer within the time limit (20 seconds), then the maximum time of 20 seconds was recorded. The delayed Post-test time to solution thus ranged from 0 to 400 seconds.

6.1.4 Statistical Analysis

To test the hypothesis and check the group difference between the non-alternating and the alternating condition, the following statistical analyses were used in the current study.

Prior to conducting substantive analyses, group equivalence between conditions in key demographics and prior calculation ability was checked. Independent-samples *t*-test was used to check the equivalence in participants' prior calculation abilities and their handedness scores. Fisher's exact test (Upton, 1992) was used to check the equivalence in gender proportions.

The analysis of covariance (ANCOVA) was applied to check the group difference on all indicators representing participants' learning performance, using the prior calculation ability as the covariate. Repeated measures ANCOVA (Algina, 1982) was conducted to further analyse the change in participants' mastery level between the immediate post-test and the delayed post-test. ANCOVA was also used to check the group difference on some cognitive load indicators, depending on their correlation with the prior calculation ability. For those cognitive load indicators that do not correlate with the prior calculation ability, the independent-samples *t*-test was used to check the group difference. Liu and Wang's (2021) *t*-test, correcting for ceiling effects and floor effects, was applied to further check the results when a potential ceiling effect or floor effect was suggested. Under previous analyses, there are assumptions about the normality of the sample and the homogeneity of variance. The normality was checked via the Shapiro-Wilk (1965) test or the Lilliefors (1967) test, depending on the sample size. The homogeneity of variances was checked via Levene's (1960) test.

The majority of statistical analyses were conducted via jamovi (2.6.26) (The jamovi project, 2024) and assumptions were checked via Intellectus Statistics (The Intellectus Statistics Team, 2021). Tests of statistical significance controlled the Type 1 error rate at 0.05.

6.2 Results

6.2.1 Can We Consider Two Groups of Participants as Equivalent?

In this experiment, participants were randomly assigned to conditions, with the expectation that the participants in two conditions would be equivalent in terms of prior calculation ability.

An independent-samples *t*-test was conducted to compare participants' prior calculation abilities. As introduced in the earlier section, the pre-test score was normally distributed in both conditions, meeting the normality assumption of the *t*-test. The assumption of homogeneity of variance was also satisfied, $F(1, 93) = 0.23$, $p = .636$. Analysis showed that two conditions did not differ on pre-test scores, $t\ (93) = -1.72$, $p = .088$, $d = -0.35$, 95% CI [-0.76, 0.06] (alternating condition: $M = 61.06$, $SD = 13.50$; non-alternating condition: $M = 56.22$, $SD = 13.85$).

A similar analysis was conducted on participants' handedness scores. The assumption of homogeneity of variance was satisfied, $F(1, 93) = 2.75$, $p = .101$. Though neither the non-alternating group (Shapiro-Wilk $W = 0.71$, $p < .001$) nor the alternating group (Shapiro-Wilk $W = 0.50$, $p < .001$) met the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). Participants' handedness scores were found to be equivalent between the non-alternating ($M = 0.88$, $SD = 0.18$) and alternating condition ($M = 0.83$, $SD = 0.37$), $t\ (93) = 0.75$, $p = .456$, $d = 0.15$, 95% CI [-0.25, -0.58].

Fisher's exact test (Upton, 1992) was used to analyse participants' gender distribution across conditions because of its suitability for small sample sizes, instead of the Pearson chi-square test (Pearson, 1900) that requires every cell's frequencies to be bigger than five (Kim, 2016). In the current study, there was an option of "prefer not to say" that was selected by only one participant, which did not meet the basic requirements of the chi-squared test. Fisher's exact test indicated that the non-alternating condition and the alternating condition were also equivalent in terms of the distribution of gender, $p = .375$.

6.2.2 Learning Performance

Did Participants' Prior Abilities Affect Their Learning Performance, Motivation, and Cognitive Load? Though participants' prior calculation ability knowledge can be considered equivalent at the group level, a correlational analysis was conducted to further check if students' prior calculation ability influenced participants' learning performance at the individual level. As Table 4 shows, participants' pre-test scores were significantly correlated with all three types of learning performance. Thus, the pre-test scores were considered as a covariate in the following analyses.

Table 4

Correlations Between Prior Ability and Learning Performance in Experiment 1

Variable	n	M	SD	1	2	3	4	5	6	7
1. Pre-test Score (/180)	95	58.77	13.81	—						
2. Practice Score (/4)	95	2.80	1.17	$.32^{**}$	—					
3. Practice Time to Solution (/120)	95	87.85	21.86	$-.47^{***}$	$-.81^{***}$	—				
4. Immediate Post-Test Score (/20)	95	12.35	5.36	$.52^{***}$	$.52^{***}$	$-.49^{***}$	—			
5. Immediate Post-Test Time to Solution (/400)	95	345.77	37.82	$-.59^{***}$	$-.50^{***}$	$.60^{***}$	$-.82^{***}$	—		
6. Delayed Post-Test Score (/20)	95	16.36	3.53	$.48^{***}$	$.43^{***}$	$-.38^{***}$	$.64^{***}$	$-.50^{***}$	—	
7. Delayed Post-Test Time to Solution (/400)	95	305.01	43.12	$-.57^{***}$	$-.47^{***}$	$.56^{***}$	$-.71^{***}$	$.84^{***}$	$-.72^{***}$	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

How Did Participants Perform in the Practice Phase? Since participants' prior abilities were found to correlate with their learning performance, the ANCOVA test (Fox et al., 2024) was used to compare participants' learning performance in the practice phase across the two conditions. Statistical analyses controlled the Type 1 error rate at 0.05. Two indicators, practice score and practice time to solution, were separately analysed with the same data processing procedure.

For the practice score, homogeneity of variances was found via Levene's (1960) test, $F(1, 93) = 1.93$, $p = .168$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.11$, $p = .006$). This test is a variation of the widely used Kolmogorov-Smirnov test with a wider range of applications and better performance (Öztuna et al., 2006; Razali & Wah, 2011), being suitable for those analyses with a sample size larger than 50 (Mishra et al., 2019); this test does not specify the expected value and variance of the distribution. As this experiment had 95 participants in total and ANCOVA tested normality based on the entire sample, the Lilliefors (1967) test was used to assess the assumption of normality in ANCOVA analysis in the current research. Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, given Olejnik and Algina's (1984) findings that ANCOVA is robust when either the assumption of normality or the assumption of homoscedasticity is violated in isolation. The interaction effect between condition and prior abilities was not statistically significant, $F(1, 91) = 0.16$, $p = .687$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. As expected, the main effect of prior ability was statistically significant, $F(1, 92) = 14.70$, $p < .001$, $\eta^2 p = .14$; participants with higher pre-test scores also gained higher practice scores. The main effect of the condition was also statistically significant, $F(1, 92) = 7.50$, $p = .007$, $\eta^2 p = .08$; participants in the non-alternating condition (estimated marginal $M = 3.12$, $SE = .16$) solved more problems correctly than those in the alternating condition (estimated marginal $M = 2.51$, $SE = .15$).

For practice time to solution, homogeneity of variances was found via Levene's (1960) test, $F(1, 93) = 3.49$, $p = .065$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.05$, $p = .817$). The interaction effect between the condition and the prior abilities was not statistically significant, $F(1, 91) = 2.07$, $p = .153$, showing that the assumption of homogeneity of regression slopes was met. Similar results on the main effects were found. The main effect of prior ability was statistically significant, $F(1, 92) =32.98$, $p < .001$, $\eta^2 p = .26$; participants with higher pre-test scores spent less time finishing practice questions. The main effect of condition was also found, $F(1, 92) = 9.03$, $p = .003$, $\eta^2 p = .09$; participants in the non-alternating condition (estimated marginal $M = 81.69$, $SE = 2.80$) spent less time finishing practice questions than those in the alternating condition (estimated marginal $M = 93.39$, $SE = 2.66$).

Thus, contrary to *H1*, both indicators showed that participants in the non-alternating condition performed better than those in the alternating condition on the practice questions.

How Did Participants Perform in the Immediate Post-Test Phase? The ANCOVA test (Fox et al., 2024) was used to compare participants' learning performance in the immediate post-test in two conditions. Analyses controlled the Type 1 error rate at 0.05. Two indicators, immediate post-test score and immediate post-test time to solution, were separately analysed with the same data processing procedure.

For the immediate post-test score, homogeneity of variances was found via the Levene (1960) test, $F(1, 93) = 0.27$, $p = .608$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.08$, $p = .168$). No interaction effect was found between the condition and the prior abilities on the immediate post-test scores, $F(1, 91) = 0.09$, $p = .771$, $\eta^2p = .00$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for immediate post-test scores, $F(1, 92) = 39.66$, $p < .001$, $\eta^2p = .30$; participants with higher pre-test scores also gained higher immediate post-test scores. The main effect of condition was statistically significant on the immediate post-test scores, $F(1, 92) = 4.65$, $p = .034$, $\eta^2p = .05$; participants in the non-alternating condition gained higher scores (estimated marginal $M = 13.42$, $SE = 0.68$) than those in the alternating condition (estimated marginal $M = 11.39$, $SE = 0.64$).

For immediate post-test time to solution, homogeneity of variances was found via the Levene (1960) test, $F(1, 93) = 0.46$, $p = .500$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.09$, $p = .057$). No interaction effect was found between the condition and the prior abilities on the immediate post-test time to solution, $F(1, 91) = 1.14$, $p = .289$, $\eta^2p = .01$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for immediate post-test time to solution, $F(1, 92) = 58.82$, $p < .001$, $\eta^2p = .39$, showing that participants with higher pre-test scores spent less time finishing immediate post-test questions. The main effect of the condition on immediate post-test time to solution was also statistically significant, $F(1, 92) = 9.03$, $p = .003$, $\eta^2 p = .09$, the participants in the non-alternating condition spent less time finishing immediate post-test questions (estimated marginal $M = 336.00$, $SE = 4.45$) than those in the alternating condition (estimated marginal $M = 354.56$, $SE = 4.21$).

Thus, contrary to *H6*, both indicators showed that participants in the non-alternating condition performed better than those in the alternating condition on the immediate post-test questions.

How Did Participants Perform in the Delayed Post-Test Phase? Procedures to analyse participants' learning performance in the delayed post-test phase were similar.

For the delayed post-test score, homogeneity of variances was found via the Levene (1960) test, $F(1, 93) = 0.62$, $p = .432$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.11$, $p = .009$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. No interaction effect was found between the condition and the prior abilities on the delayed post-test scores, $F(1, 91) = 0.03$, $p = .863$, $\eta^2 p = .00$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for delayed post-test scores, $F(1, 92) = 27.28$, $p < .001$, $\eta^2 p = .23$; participants with higher pre-test scores also gained higher delayed post-test scores. The main effect of the condition was not significant for delayed post-test scores, $F(1, 92) = 0.11$, $p = .742$, $\eta^2 p = .00$; participants in the non-alternating condition gained similar scores (estimated marginal $M =$

16.47, *SE* = 0.47) to those in the alternating condition (estimated marginal *M* = 16.26, *SE* = 0.45).

Since the mean delayed post-test scores were high in both conditions, potential ceiling effects may exist in the delayed post-test scores. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For the delayed post-test score, participant score of 20, the highest possible score, constituted 11.1% and 14.0% for non-alternating and alternating conditions respectively. If the criteria were relaxed to the highest two scores, participant scores of 19 or 20 constituted 28.9% and 44% of scores for the non-alternating and alternating conditions respectively. Based on McHorney and Tarlov's (1995) guidance that ceiling or floor effects are likely to be an issue when more than 15% of respondents respond with either the highest or lowest possible score, a ceiling effect was evident for the delayed post-test score. Meier and Feeley's (2022) criterion based on the standard deviation also suggested the presence of a ceiling effect, as the sum of the mean plus 1-2 standard deviations equalled or exceeded the maximum test score. For the delayed post-test score, the sum of the mean plus 1 *SD* was 19.89, and the sum of the mean plus 2 *SD* was 23.42 (i.e., substantially higher than the maximum possible score), further confirming the existence of a ceiling effect.

Considering the ceiling effect was detected on the delayed post-test score, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the result against this potential assumption violation. Under this analysis, $t(93) = -0.69$, $p = .521$, $d = -0.15$, 95% CI [-2.58, 1.33], indicating that there was indeed no significant difference in participants' delayed post-test scores between the non-alternating finger tracing group (corrected *M* = 16.52, *SD* = 3.97) and the alternating fingers tracing group (corrected *M* = 17.15, *SD* = 4.52).

For the delayed post-test time to solution, homogeneity of variances was found via the Levene (1960) test, $F(1, 93) = 0.57$, $p = .452$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.08$, $p = .143$). No interaction effect was found between the condition and the prior abilities on the delayed post-test time to solution, $F(1, 91) = 0.20$, $p = .656$, $\eta^2p = .00$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for delayed post-test time to solution, $F(1, 92) = 49.87$, $p < .001$, $\eta^2p = .35$, showing that participants with higher pre-test scores spent less time finishing delayed post-test questions. The main effect of condition was found on delayed post-test time to solution, $F(1, 92) =4.26$, $p = .042$, $\eta^2p = .04$; participants in the non-alternating condition (estimated marginal *M* = 297.07, *SE* = 5.26) solved problems more quickly than those in the alternating condition (estimated marginal *M* = 312.15, *SE* = 4.99).

Thus, contrary to *H7*, the delayed post-test score showed that participants in the non-alternating condition performed similarly to those in the alternating condition, and the delayed post-test time to solution showed that participants in the non-alternating condition performed better than those in the alternating condition on the immediate post-test questions.

How Did Participants' Learning Performance Change Between the Immediate Post-Test and the Delayed Post-Test? To further analyse the change of participants' mastery level between the immediate post-test and the delayed post-test, a repeated measures ANCOVA (Algina, 1982) was conducted. The pre-test score worked as the covariate because it was correlated with participants' immediate post-test performance and delayed post-test performance. Two indicators, the test score and the time to solution, were separately analysed with the same data processing procedure.

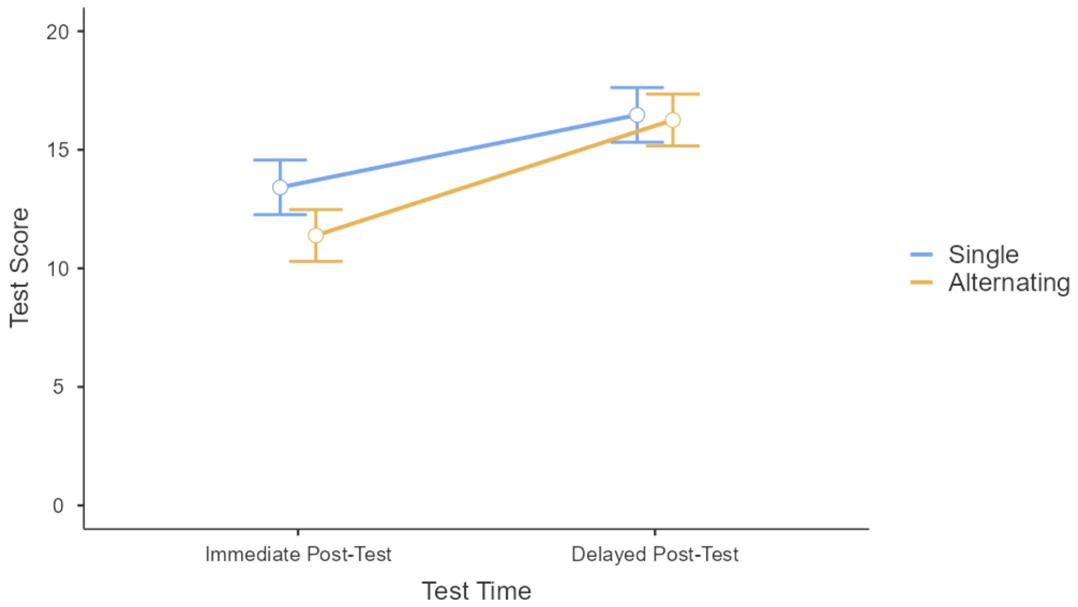
For the test score, homogeneity of variances was found via the Levene (1960) test both in the immediate post-test, $F(1, 93) = 0.27$, $p = .608$, and the delayed post-test, $F(1, 93) = 0.62$, $p = .432$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.07$, $p = .020$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

The main effect for the test time was significant, $F(1, 92) = 96.54$, $p < .001$, $\eta^2p = .51$, indicating there were significant differences between the immediate post-test score and delayed post-test score; participants gained higher test scores in the delayed post-test (estimated marginal $M = 16.36$, $SE = 0.32$) than in the immediate post-test (estimated marginal $M = 12.40$, $SE = 0.46$). The main effect for the condition was not significant, $F(1, 92) = 2.58$, $p = .112$, $\eta^2p = .03$, indicating participants in the non-alternating condition gained similar test scores (estimated marginal $M = 14.94$, $SE = 0.50$) to those in the alternating condition (estimated marginal $M = 13.82$, $SE = 0.48$). The main effect for pre-test score was significant, $F(1, 92) = 44.65$, $p < .001$, $\eta^2p = .33$, indicating participants with higher prior calculation ability also gained higher test scores. The interaction effect between test time and condition was significant, $F(1, 92) = 4.90$, $p = .029$, $\eta^2p = .09$, indicating that participants allocated to different conditions had different trends of change on their test scores across the

24-hour delayed time. As Figure 6 shows, participants in the alternating condition showed greater progress during the 24-hour delayed time than those in the non-alternating condition.

Figure 6

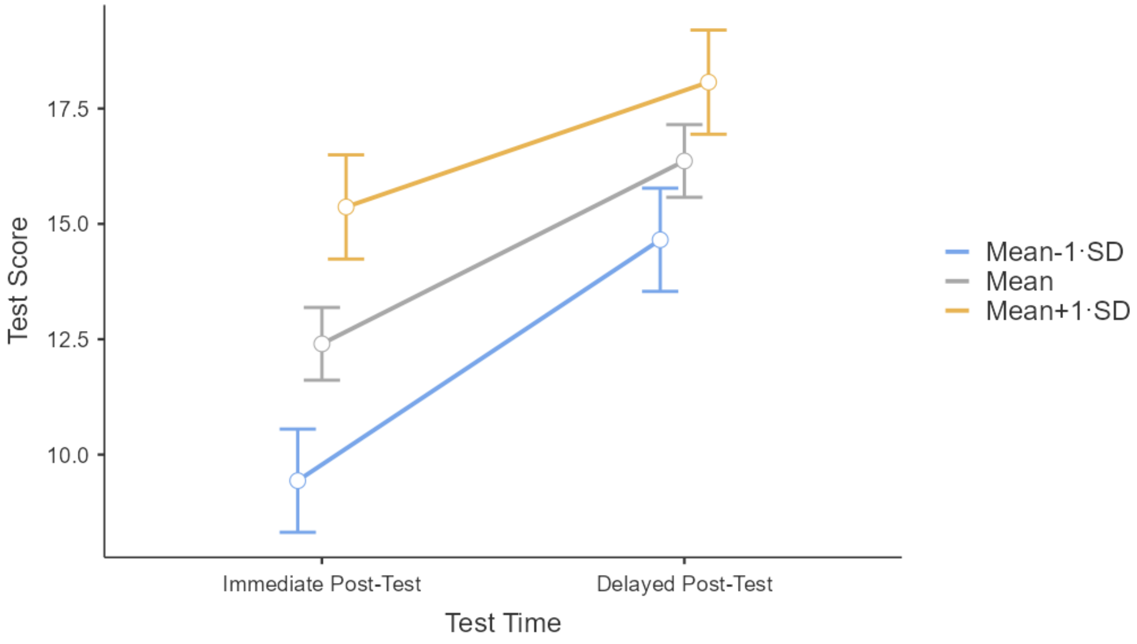
Participants' Test Scores in Two Conditions in Immediate Post-Test and Delayed Post-Test in Experiment 1



The interaction effect between the test time and pre-test scores was significant, $F(1, 92) = 9.40$, $p = .003$, $\eta^2p = .05$, indicating that participants with different prior calculation ability may have had different trends of change in their test scores during the 24-hour delayed time. To further check the specific "turning point" of the interaction effect, a Johnson-Neyman analysis (Johnson & Fay, 1950; Johnson & Neyman, 1936) was applied to pick the exact point (Hayes & Matthes, 2009). Johnson-Neyman analysis treats each value of the moderator as a possible "turning point" where the effect of the focal predictor may shift from non-significance to significance (Krishna, 2016), identifying a specific range of the pre-test score where the effect of test time is significant and non-significant in current research (Hayes & Matthes, 2009). This analysis showed that the effect of test time was only significant when participants' pre-test total score was lower than 83.78, covering 97.89% (93 participants) of total observations (95 participants). As Figure 7 shows, participants with lower pre-test total scores gained more progress in the delayed post-test. Combining the results of the Johnson-Neyman analysis, we can conclude that participants whose pre-test total score was lower than 83.78 gained significantly higher test scores in the delayed post-test compared to the immediate post-test.

Figure 7

Participants' Test Scores Under Similar Pre-test Scores in Two Tests in Experiment 1



For the time to solution, homogeneity of variances was found via the Levene (1960) test both in the immediate post-test, $F(1, 93) = 0.46$, $p = .500$, and the delayed post-test, $F(1, 93) = 0.57$, $p = .452$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.07$, $p = .032$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

The main effect for the test time was significant, $F(1, 92) = 274.78$, $p < .001$, $\eta^2p = .75$, indicating there were significant differences between the immediate post-test time to solution and delayed post-test time to solution; participants spent less time completing the test questions in the delayed post-test (estimated marginal $M = 304.61$, $SE = 3.60$) than in the immediate post-test (estimated marginal *M* = 345.28, *SE* = 3.04). The main effect for the condition was significant, $F(1, 92) = 7.15$, $p = .009$, $\eta^2p = .07$, indicating participants in the non-alternating condition spent less time completing the test questions (estimated marginal *M* = 316.54, *SE* = 4.53) than those in the alternating condition (estimated marginal *M* = 333.35, *SE* = 4.29). The main effect for pre-test score was significant, $F(1, 92) = 61.90$, $p < .001$, $\eta^2p = .40$, indicating participants with higher prior calculation ability spent less time completing the test questions.

The interaction effect between test time and condition was not significant, $F(1, 92) = 0.49$, $p = .487$, $\eta^2p = .01$, indicating that participants allocated to different conditions had similar trends of change on their time to solution across the 24-hour delayed time. The interaction effect between the test time and pre-test scores was also not significant, $F(1, 92) = 0.73$, $p = .395$, $\eta^2p = .01$, indicating that participants with different prior calculation ability had similar trends of change in their time to solution during the 24-hour delayed time.

6.2.3 Post-Lesson Self-Report

Did Participants' Prior Abilities Affect Their Learning Performance, Motivation, and Cognitive Load? As for learning performance, a correlational analysis was also conducted to determine prior ability's relation with motivation and cognitive load. Table 5 presents detailed results of this analysis. Intrinsic cognitive load was significantly correlated with pre-test scores, which should be considered in the following analysis.

Table 5

Correlation Between Prior Ability, Motivation, and Cognitive Load in Experiment 1

Variable	n	M	*SD*	1	2	3	4	5
1. Pre-Test Score (/180)	95	58.77	13.81	—				
2. Intrinsic Motivation (/8)	95	5.66	1.54	.01	—			
3. Intrinsic Cognitive Load (/8)	95	1.07	1.05	-.21*	-.16	—		
4. Extraneous Cognitive Load (/8)	95	1.69	1.37	-.03	-.14	.55***	—	
5. Germane Processing (/8)	95	6.27	1.12	-.08	.30**	-.46***	-.52***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

What is the Intrinsic Motivation Reported by Participants? Since correlational analysis showed there was no clear correlation between participants' prior knowledge and their intrinsic motivation, an independent-samples *t*-test was applied to analyse participants' self-reported intrinsic motivation.

The assumption of normality was met in the alternating group (Shapiro-Wilk $W = 0.97$, $p = .229$), but not in the non-alternating group (Shapiro-Wilk $W = 0.93$, $p = .009$). The assumption of homogeneity of variance was satisfied, $F(1, 93) = 0.43$, $p = .513$. Though one of the conditions did not meet the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). The difference in self-reported intrinsic motivation ratings reported by the non-alternating condition ($M = 5.57$, $SD = 1.66$) and the alternating condition ($M = 5.74$, $SD = 1.45$) was not statistically significant, $t(93) = -0.54$, $p = .592$, $d = -0.11$, 95% CI [-0.51, 0.29].

Considering the ceiling effect was detected on the average intrinsic motivation scores, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the result against this potential assumption violation. Under this analysis, $t(93) = -0.53$, $p = .602$, $d = -0.11$, 95% CI [-0.90, 0.53], indicating that there was indeed no significant difference in participants' self-reported intrinsic motivation between the non-alternating finger tracing group (corrected $M = 5.61$, $SD = 1.73$) and the alternating fingers tracing group (corrected $M = 5.79$, $SD = 1.54$).

Overall, participants in both groups reported relatively high levels of intrinsic motivation. However, contrary to *H2*, there was no significant difference in their motivation during the lesson between groups.

What is the Intrinsic Cognitive Load Reported by Participants? Since participants' prior abilities were found to correlate with their intrinsic cognitive load, the ANCOVA test was used to compare participants' intrinsic cognitive load in two conditions.

Homogeneity of variances was found via the Levene (1960) test, $F(1, 93) = 0.94$, $p = .336$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.10$, $p = .034$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. The interaction effect between condition and students' prior calculation ability was found for intrinsic cognitive load, $F(1, 91) = 5.91$, $p = .017$, $\eta^2 p = .06$, which indicated that the assumption of homogeneity of regression slopes was not met. The interaction factor was therefore kept in the ANCOVA model.

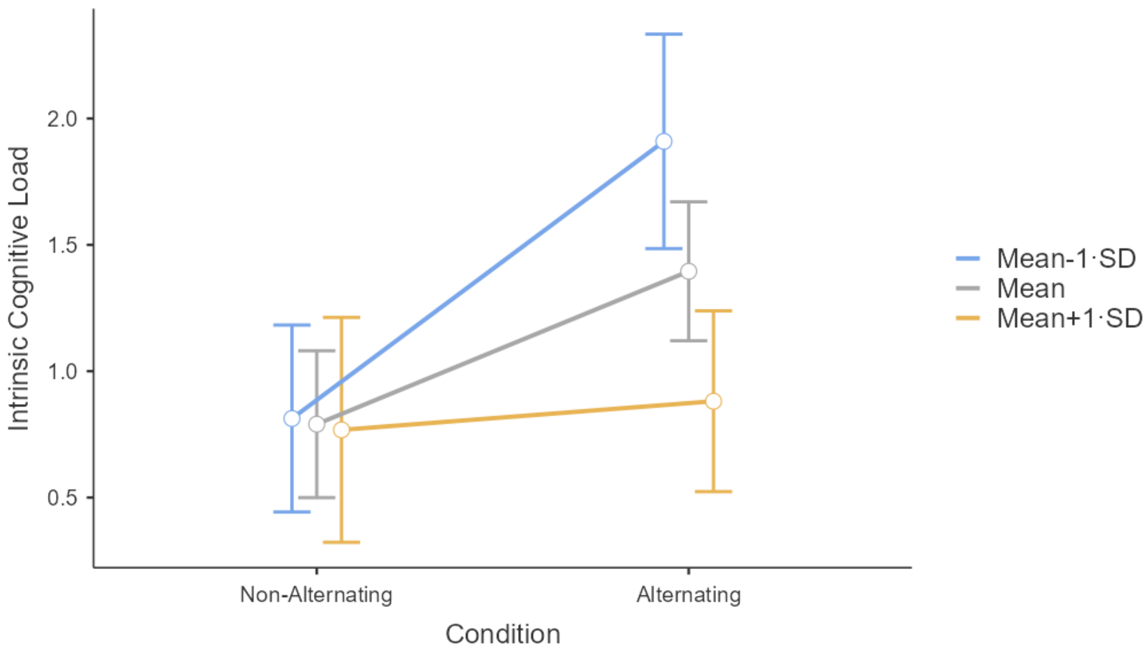
The main effect of prior ability was statistically significant, $F(1, 91) = 7.04$, $p = .009$, $\eta^2p = .07$; participants with higher pre-test scores perceived lower intrinsic cognitive load. The main effect of condition was also statistically significant, $F(1, 91) = 9.37$, $p = .003$, $\eta^2p = .09$; participants in the non-alternating condition (estimated marginal $M = 0.79$, $SE = .15$) perceived lower intrinsic cognitive load than those in the alternating condition (estimated marginal $M = 1.40$, $SE = .14$).

As noted above, the interaction effect between condition and prior calculation ability was significant, indicating that participants with different prior calculation abilities may be affected by tracing strategies to varying degrees. To further check the specific “turning point” of the interaction effect, a Johnson-Neyman analysis (Johnson & Fay, 1950; Johnson & Neyman, 1936) was applied to pick the exact point (Hayes & Matthes, 2009). Johnson-Neyman analysis treats each value of the moderator as a possible “turning point” where the effect of the focal predictor may shift from non-significance to significance (Krishna, 2016), identifying a specific range of the pre-test score where the effect of the condition is significant and non-significant in current research (Hayes & Matthes, 2009). This analysis showed the effect of the condition was only significant when participants’ pre-test total score was lower than 63.75, covering 63.16% (60 participants) of total observations. As Figure 8 shows, participants with lower pre-test total scores reported much higher intrinsic cognitive load in the alternating condition. Combining the results of the Johnson-Neyman analysis, we can conclude that participants whose pre-test total score was lower than 63.75 perceived significantly lower intrinsic cognitive load when they were allocated to a non-alternating condition compared with the alternating condition.

Overall, participants in both groups reported relatively low levels of intrinsic cognitive load. However, contrary to *H3*, an interaction effect was found, suggesting participants with lower prior calculation ability perceived significantly lower intrinsic cognitive load when they were allocated to a non-alternating condition compared with the alternating condition.

Figure 8

Participants' Intrinsic Cognitive Load Under Similar Pre-test Scores in Two Conditions in Experiment 1



What is the Extraneous Cognitive Load Reported by Participants? Since correlational analysis showed there was no correlation between participants' prior knowledge and their extraneous cognitive load, an independent-samples *t*-test was applied to investigate group differences in participants' extraneous cognitive load.

The assumption of homogeneity of variance was satisfied, $F(1, 93) = 0.56$, $p = .455$. Though neither the non-alternating group (Shapiro-Wilk $W = 0.87$, $p < .001$) nor the alternating group (Shapiro-Wilk $W = 0.94$, $p = .018$) met the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). The difference in self-reported extraneous cognitive load ratings reported by the alternating condition ($M = 1.45$, $SD = 1.35$) and the non-alternating condition ($M = 1.90$, $SD = 1.37$) was not statistically significant, $t(93) = -1.62$, $p = .109$, $d = -0.33$, 95% CI [-0.74, 0.07]. Given the apparent floor effect for this variate, Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects was used to check the result against this potential assumption violation. Under this analysis, $t(93) = -2.62$, $p = .013$, $d = -0.60$, 95% CI [-0.82, -0.10], indicating participants in the alternating group (corrected $M = 1.95$, $SD = 0.69$) perceived higher extraneous cognitive load than those in the non-alternating group (corrected $M = 1.49$, $SD = 0.85$).

Overall, participants in both groups reported relatively low levels of extraneous cognitive load, with a floor effect apparent in participants' reports of extraneous cognitive load. Contrary to *H4,* the average extraneous cognitive load between the non-alternating finger tracing group and the alternating finger tracing group showed a significant difference after considering the floor effect, suggesting participants in the non-alternating finger tracing group perceived lower extraneous cognitive load.

What is the Germane Processing Reported by Participants? Since correlational analysis showed there was no clear correlation between participants' prior knowledge and their germane processing scores, an independent-samples *t*-test was applied to investigate group differences in participants' reports of germane processing.

The assumption of normality was met in the non-alternating group (Shapiro-Wilk W = 0.95, p = .061), but not in the alternating group (Shapiro-Wilk W = 0.95, p = .024). The assumption of homogeneity of variance was satisfied, $F(1, 93) = 0.49$, p = .486. Though one of the conditions did not meet the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). The difference in self-reported intrinsic motivation ratings reported by the non-alternating condition (M = 6.37, SD = 1.09) and the alternating condition (M = 6.18, SD = 1.16) was not statistically significant, $t(93) = 0.83$, p = .409, d = 0.17, 95% CI [-0.23, 0.57].

Considering the ceiling effect was detected on the average germane processing scores, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the data to check the result against this potential assumption violation. Under this analysis, $t(93) = 0.78$, p = .441, d = 0.16, 95% CI [-0.31, 0.70], indicating that there was indeed no significant difference in participants' self-reported germane processing between the non-alternating finger tracing (corrected M = 6.40, SD = 1.16) group and the alternating fingers tracing group (corrected M = 6.21, SD = 1.22).

Overall, participants in both groups reported relatively high levels of germane processing. However, contrary to *H5*, there was no significant difference in their germane processing scores between groups.

6.3 Discussion

6.3.1 How Did Tracing Conditions Affect Participants' Learning Performance?

Previous analysis showed that participants in the non-alternating condition performed better than those in the alternating condition, indicating that the non-alternating finger tracing strategy may enhance learning better than the alternating fingers tracing strategy. Lesson materials emphasising alternating tracing were designed to further support participants' learning, but the current results did not support this hypothesis. To consider why the results did not align with expectations, we further summarise the current findings.

Descriptively, the participants' accuracy rates (see Table 6) exhibited a "✓" shaped curve across the three phases. Participants in both groups had moderate accuracy in the practice phase, lower accuracy in the immediate post-test phase, and highest accuracy in the delayed post-test phase. Repeated-measure analysis also suggested that participants' performance in the delayed post-test was significantly better than in the immediate post-test. Better performance in the delayed post-test suggested the possible presence of a practice effect (Anastasi, 1988; Campbell & Stanley, 1963), especially considering that the participants completed identical questions in the immediate post-test and the delayed post-test.

Comparing the two conditions, participants in the non-alternating condition gained higher scores in practice questions and immediate post-test questions, but this advantage was not observed in the delayed post-test questions. Repeated-measure analysis suggested that participants in the alternating condition made significantly greater progress during the 24-

hour delayed time than those in the non-alternating condition. The aforementioned trends suggested that the superior performance of the non-alternating group over the alternating group gradually diminished, which finally resulted in a similar performance between the two groups in the delayed post-test. There were two possible explanations for this diminishing phenomenon: a) The effects of the tracing interventions were no longer effective after 24 hours, so no difference in learning performance was observed in the delayed post-test between the two groups; b) The benefits of alternating tracing may take longer to manifest, potentially explaining why participants in the alternating group performed worse at the beginning but caught up in the delayed post-test. However, considering that participants' performance in the delayed post-test became better instead of worse, the explanation that alternating fingers tracing's effect may require more time to become apparent is arguably more reasonable. Though we cannot draw a definite conclusion, we can reasonably imagine participants in the alternating group may perform significantly better than those in the non-alternating group in a much-delayed post-test (e.g. 48 hours or 72 hours later) from the trends visible so far.

Table 6

Participants' Accuracy Rates in Three Tests in Experiment 1

Variable	Total	Non-Alternating	Alternating
Practice Questions (/4)			
Mean Score	2.80	3.04	2.58
Accuracy Rate	70.00%	76.10%	64.50%
Immediate Post-Test (/20)			
Mean Score	12.35	12.87	11.88
Accuracy Rate	61.74%	64.34%	59.40%
Delayed Post-Test (/20)			
Mean Score	16.36	16.16	16.54
Accuracy Rate	81.79%	80.78%	82.70%

6.3.2 How Did Tracing Conditions Affect Participants' Intrinsic Motivation and Cognitive Load?

Previous analysis described participants' intrinsic motivation and cognitive load in different tracing conditions in detail. Participants in the non-alternating condition had relatively high levels of intrinsic motivation and germane processing, and they were similar to those in the alternating condition. Participants in both groups perceived relatively low levels of intrinsic cognitive load, and an interaction effect was found between participants' calculation abilities and the condition on the intrinsic cognitive load. Participants whose pre-test total score was lower than 63.8 perceived significantly lower intrinsic cognitive load when they were allocated to a non-alternating condition compared to the alternating condition. Participants in the non-alternating condition perceived relatively low levels of extraneous cognitive load, and lower than those in the alternating condition. These findings thus partially support the research hypothesis for intrinsic load. Supplementary analyses revealed that participants with lower pre-test scores (i.e., lower calculation ability) in the alternating fingers tracing condition perceived significantly higher intrinsic cognitive load than those in the non-alternating finger tracing condition. Participants in the alternating condition were asked to switch their index fingers during the lesson, introducing higher variability. Thus, higher perceived intrinsic cognitive load was reasonable for the participants in this group. However, this difference between groups was diminished for participants with greater prior calculation ability, where participants with high calculation ability perceived similar levels of intrinsic cognitive load across the two conditions. This finding was consistent with the basic tenet of cognitive load theory that the intrinsic cognitive load a student experiences is in part a function of her prior knowledge (Sweller, 2010; Sweller et al., 1998).

However, participants in the alternating fingers tracing group also reported higher extraneous cognitive load, which was unexpected. The initial hypothesis posited that the extraneous load should be similar between the two conditions. We propose that some unintended design-related distractions caused these results. After comparing this study's learning materials with the materials in Wang's (2022) research, we suggest that a potential split attention effect (Chandler & Sweller, 1991) may exist in the current research because a separate text box on the side of the page was used to guide participants' tracing actions, instead of placing these instructions right below the learning steps in the middle of the page. The separate text box on the side was designed to help participants discriminate the calculation-related learning steps and the required learning actions (i.e., which hand should be used to trace), to reduce participants' intrinsic cognitive load. However, additional information on the side may have required additional "search and match" eye movements (Sweller, 1989) by students in the alternating condition, resulting in higher extraneous cognitive load. As participants in the alternating tracing group were requested to trace the learning material in a more complex way, the instructions they gained from the separate text box were more complex. Thus, participants in this group may be affected by the split attention effect more deeply, resulting in a higher level of extraneous cognitive load.

No difference in germane processing was detected. We hypothesised that participants in the alternating fingers tracing condition might report higher germane processing, because the alternating action would help them clearly distinguish the learning steps and facilitate information processing. However, the unexpected split attention caused by the separate box and its guidance on the alternating action may have turned this design into a burden.

Unexpectedly higher extraneous cognitive load in the alternating condition suggested deficiencies in the current experimental material's design, so a new experiment was then designed to reduce the potential impact of split attention and subsequent extraneous cognitive load. Experiment 2 was then conducted to further explore the effects of alternating finger tracing on learning processes and outcomes.

Chapter Seven: Experiment 2

As the results of Experiment 1 suggested potential issues with the design of lesson materials, a new experiment was planned to further investigate the effect of alternating-fingers tracing on learning and cognitive load. Based on previous inferences and direct feedback from some participants after the experiment, we revised some of the experimental materials. Specifically, the tracing practice page, the presentation of the tracing introduction, the difficulty arrangement in the worked examples, and the difficulty level of the immediate and delayed post-tests were adjusted. More detailed adjustments will be described in the following sections.

7.1 Method

7.1.1 Sample Size and Participants

Power Analysis. To meet the power requirements of the current research, we kept the parameters of the power analysis as the first experiment. Again, the power analysis was carried out as a one-tailed test, with effect size $d = 0.57$, $\alpha = 0.05$ and desired power = 0.9, which gave a result that at least 54 participants for each group and 108 participants in total were required.

Participants. The second experiment was conducted in March 2024 at the Shanwei Campus of South China Normal University (SCNU), Guangdong, China. All participants were SCNU college students and teachers. A total of 112 participants attended the experiment. One participant dropped out halfway and did not attend the second day's delayed post-test, and the remaining 111 participants completed the whole experiment. All participants were educated to at least the undergraduate level (including those who were currently in undergraduate programs) and were able to read English in paragraphs. All participants were randomly assigned to the control condition (applying a non-alternating finger tracing strategy, N = 56) and the experimental condition (application of an alternating fingers tracing strategy, N = 56). The gender ratio of the participants was not balanced (89 females, 22 males, 1 preferred not to say), but it was consistent with the distribution of the population where the data were collected. Participants' ages ranged from 18 to 23 years old (M = 19.08, SD = 1.02).

7.1.2 Materials and Procedures

As in Experiment 1, recruited participants took part in the experiment individually in a quiet room with guidance from the PhD researcher (Xufei Zhang). Experiment 2 also contained a delayed post-test, which required participants to return on the next day to complete it. The paper-based experimental materials were further adjusted based on Experiment 1; the full set of experimental materials is given in Appendix C. The whole experiment needed about 30-40 minutes, with the first day lasting 25-30 minutes and the second day lasting 5-10 minutes. The first day's experiment started with the consent phase, followed by the demographics phase, the pre-test phase, the learning phase, the self-report phase, and the immediate post-test phase. The second day's experiment contained the delayed post-test phase and the debriefing phase.

Participants in both conditions received identical instructions except for the learning phase, in which they were told to use different tracing strategies (according to their condition). To avoid ambiguity caused by translation, all the materials mentioned above were paper-based and presented in printed English (adapted from materials used in Wang et al., 2022). Though English was not the first language of the participants, English as a compulsory course in the national college entrance exam had been studied by all participants for at least 10 years (from Grade 3 to Grade 12). With years of compulsory English courses, all participants were able to read English in paragraphs. Additional oral instruction between phases was given in the participants' first language (Mandarin). Participants were also encouraged to ask in Mandarin if they had any doubts about the experiment. The whole experiment was videotaped after gaining participants' consent. Recorded participants' tracing actions were counted as a check on treatment fidelity (Mowbray et al., 2003), in line with previous CLT research (Eielts et al., 2020; Galbraith & Ginns, 2023).

Consent Phase. As in Experiment 1, participants were notified of the basic information about the experiment and signed a consent form. To avoid the subject-expectancy effect (Rosenthal & Fode, 1963), the current research hid the real purpose of the experiment from participants. Participants were not informed that the experiment was designed to research the tracing strategy. Instead, they were told that this experiment was used to research a new method of two-digit multiplication.

To randomly assign participants to experimental conditions, participants with odd-numbered positions in the experiment sequence were asked to throw a die to determine their condition assignment. If the result of the die roll was an odd number, the participant was assigned to the non-alternating group; if the result of the die roll was an even number, the participant was assigned to the alternating group. Participants with even-numbered positions in the experiment sequence did not need to throw a die, and they were automatically assigned to the opposite condition of the preceding participant.

The video camera was aimed at the desktop portion and only captured participants' hand movements. Video recording commenced after the participant signed the consent form.

Demographics Phase. The experiment officially started with participants answering a survey about their basic demographic information. Their gender and age were collected by choice and fill-in-the-blank questions. A handedness survey (Nicholls et al., 2013) with ten questions, as in Experiment 1, was followed.

Pre-Test Phase. The pre-test phase was designed to know participants' prior knowledge. The calculation fluency test (Sowinski et al., 2014) as used in Experiment 1 was completed in this phase.

Learning Phase. The learning material was very similar to that used in Experiment 1, with the following adjustments. The tracing practice page was more detailed compared to the tracing practice page used in Experiment 1 (see Figure 9 and Figure 10). In both experiments, the tracing practice pages were aimed at giving participants a sense of what finger tracing is and how they should trace the key ellipses with the requested index finger. In Experiment 1, the tracing practice page only showed the tracing guidance sentence and two different kinds of ellipses, one vertical and one diagonal. However, in Experiment 2, the tracing practice page demonstrated a complete problem-solving process for a three-digit number addition, identical in format to the worked examples in the formal learning phase. The only difference was that the learning task was a two-digit number multiplication in the formal worked examples. In Experiment 1, participants in both conditions were presented with the same tracing page, on which the first instruction to practise tracing asked participants to use the writing hand's index finger, and the second instruction to practise tracing asked participants to use the non-writing hand's index finger. However, in Experiment 2, participants received the tracing practice page according to their assigned condition. Participants in the non-alternating condition received a tracing practice page that only requested them to use the writing hand's index finger to trace, while participants in the alternating condition received a tracing practice page that requested them to first use the writing hand's index finger to trace the first step and then use the non-writing hand's index finger to trace the next step. A more detailed and identical practice page was expected to help participants better understand their learning tasks before entering formal learning, thereby avoiding their confusion about the learning or tracing guidance after the formal learning phase had started.

When participants practised the tracing, they were told that they could ask any questions about the learning requirements based on the practice page, which closely resembled formal learning. They were also informed that, since the formal learning phase would be timed, they would not be able to ask the researcher any questions during this phase. The researcher double-checked with the participants if they totally understood the learning requirements and the words involved in the learning steps before they proceeded to formal learning.

Figure 9

Tracing Practice Page Used in Experiment 1

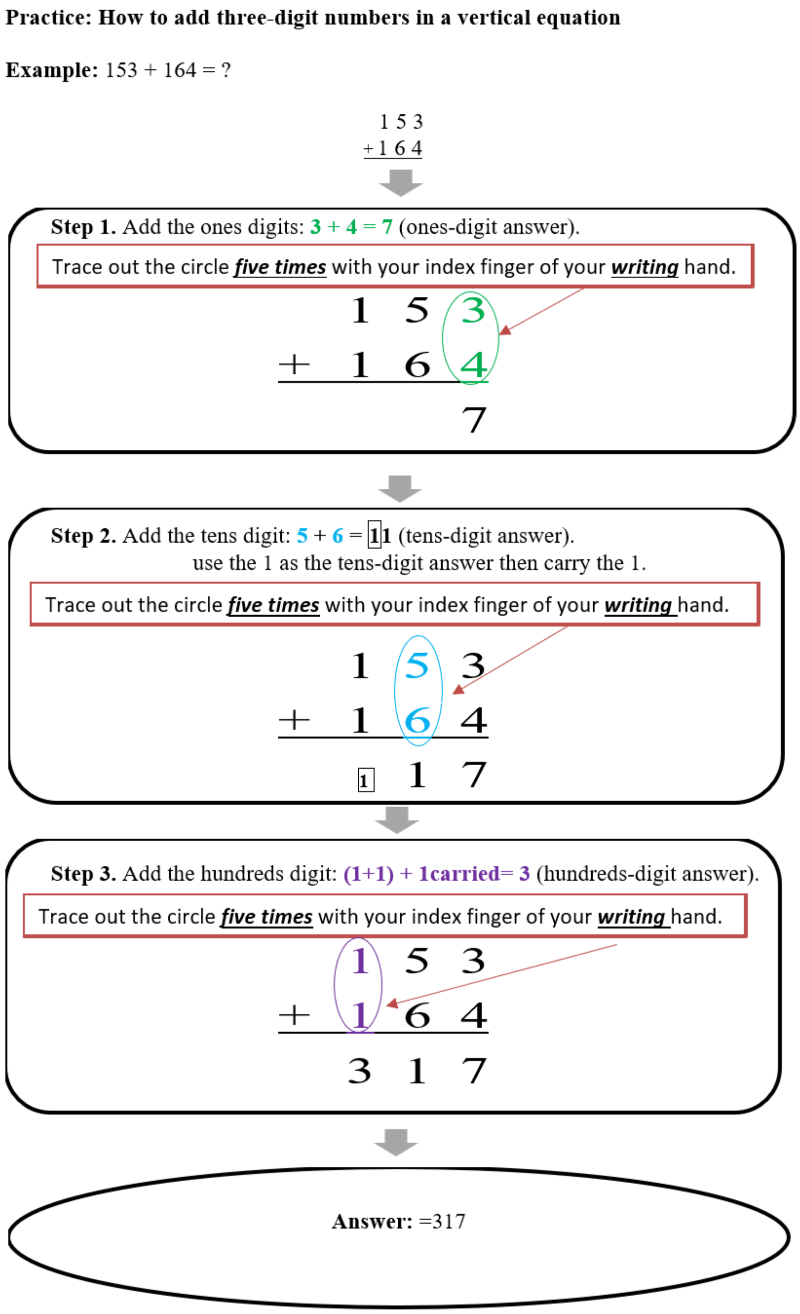
Tracing Practice

Please trace out the circle **five times** with your index finger of your **writing hand**.

Please trace out the circle **five times** with your index finger of your **non-writing hand**.

Figure 10

Tracing Practice Page Used in Experiment 2



As in Experiment 1, the eight pages of instruction included four worked examples and four practice questions. All the worked examples were about the same kind of method, but the worked examples got progressively more difficult. The worked examples were adapted from Wang et al. (2022) and Experiment 1; materials used in the present experiment further adjusted some specific numbers in the worked examples to keep them at the appropriate difficulty level. The first worked example did not contain any carry-on calculation during the solving process (see Figure 11), the second worked example contained one carry-on calculation during the solving process (see Figure 12), and the third and fourth worked examples contained two carry-on calculations in the process (see Figure 13). The fourth worked example was more complicated than the third one because its carry-on calculation in step 2 required additional calculation to determine the specific carry-on digit (see Figure 14). Based on the difficulty and complexity of the four worked examples, the current research defined these four types of two-digit number multiplication as level one, level two, level three, and level four.

As the figures showed, for each worked example, the two-digit multiplication problem was solved in three steps and instructions were provided at each step. The instruction included a short sentence explaining the step's calculating process, colour-coded horizontal and vertical formulas with key numbers in them, and colour-coded ellipses displaying the key numbers in the vertical formula. The correct answer to the whole worked example was shown at the end of the three steps.

Figure 11

A Worked Example Without any Carry-on Calculation in Experiment 2

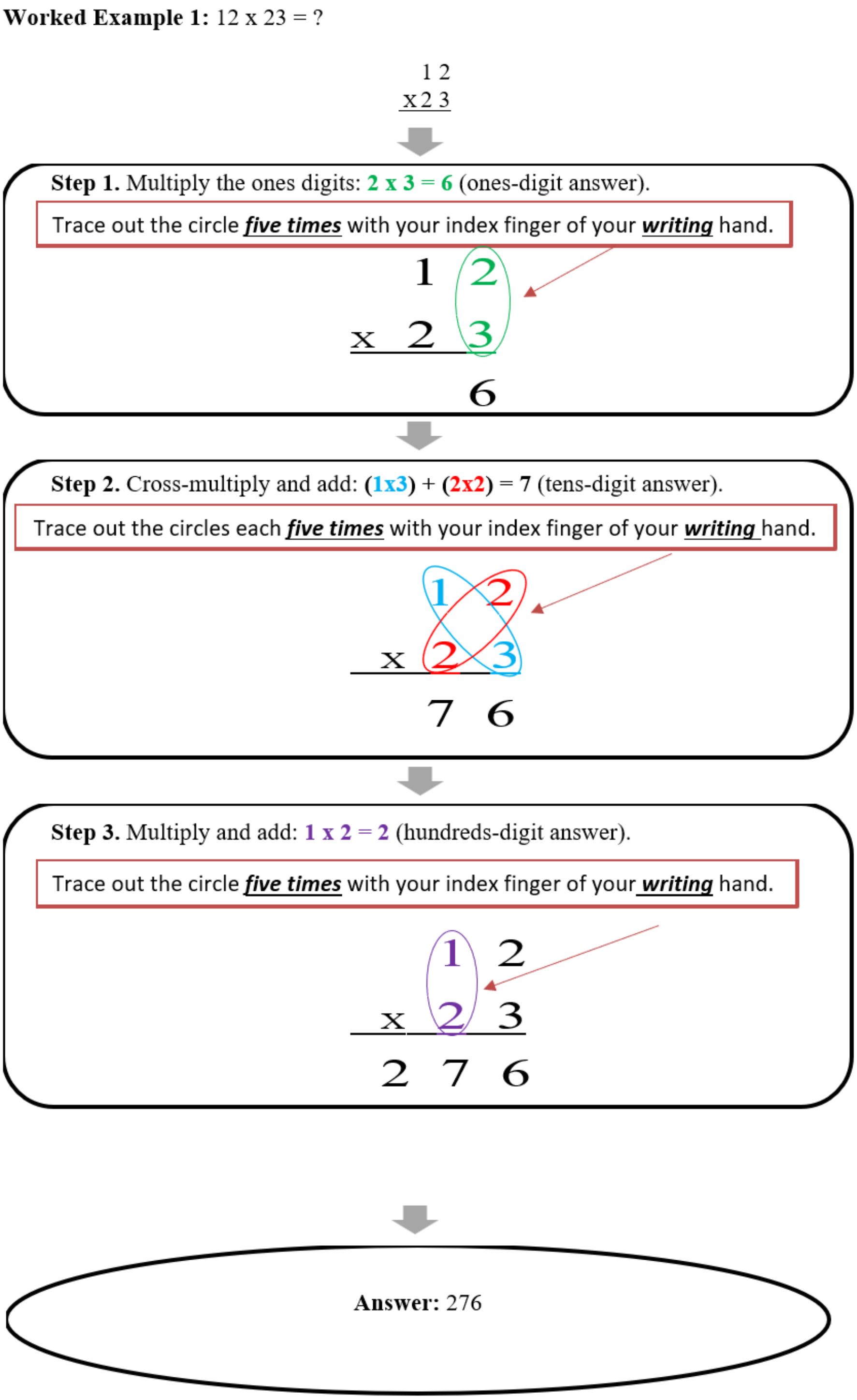


Figure 12

A Worked Example With One Carry-on Calculation in Experiment 2

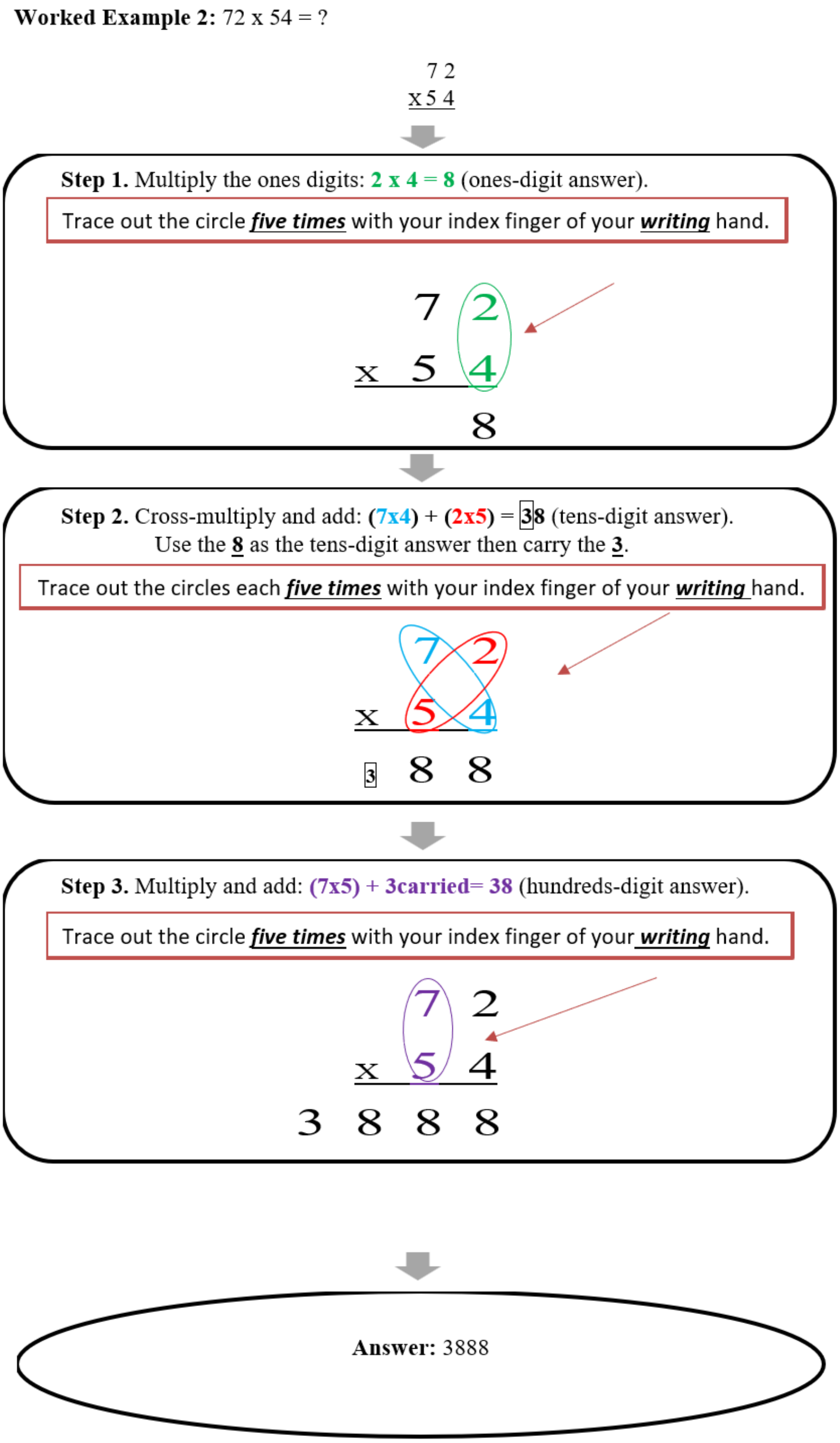


Figure 13

A Worked Example With Two Carry-on Calculations (Simple) in Experiment 2

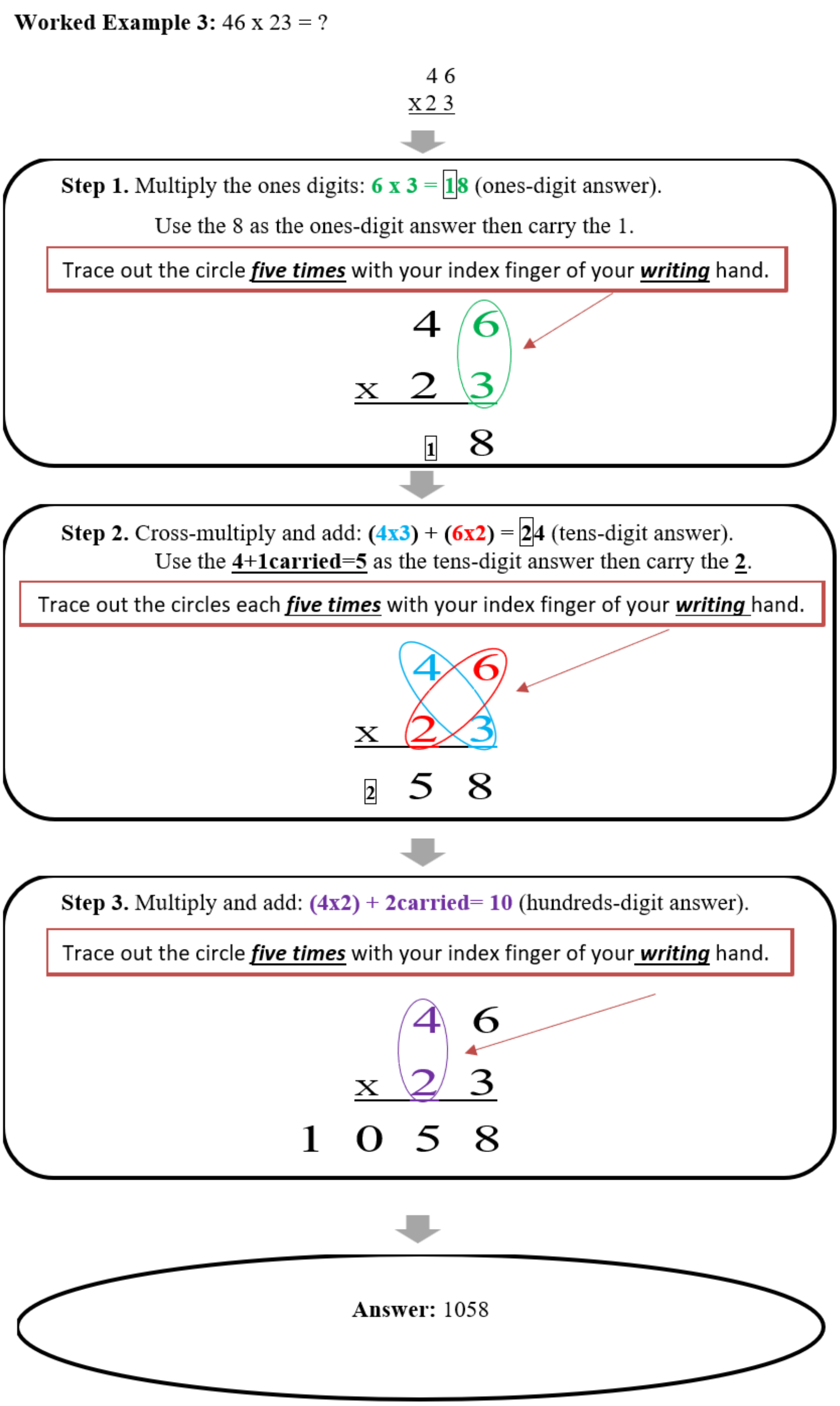
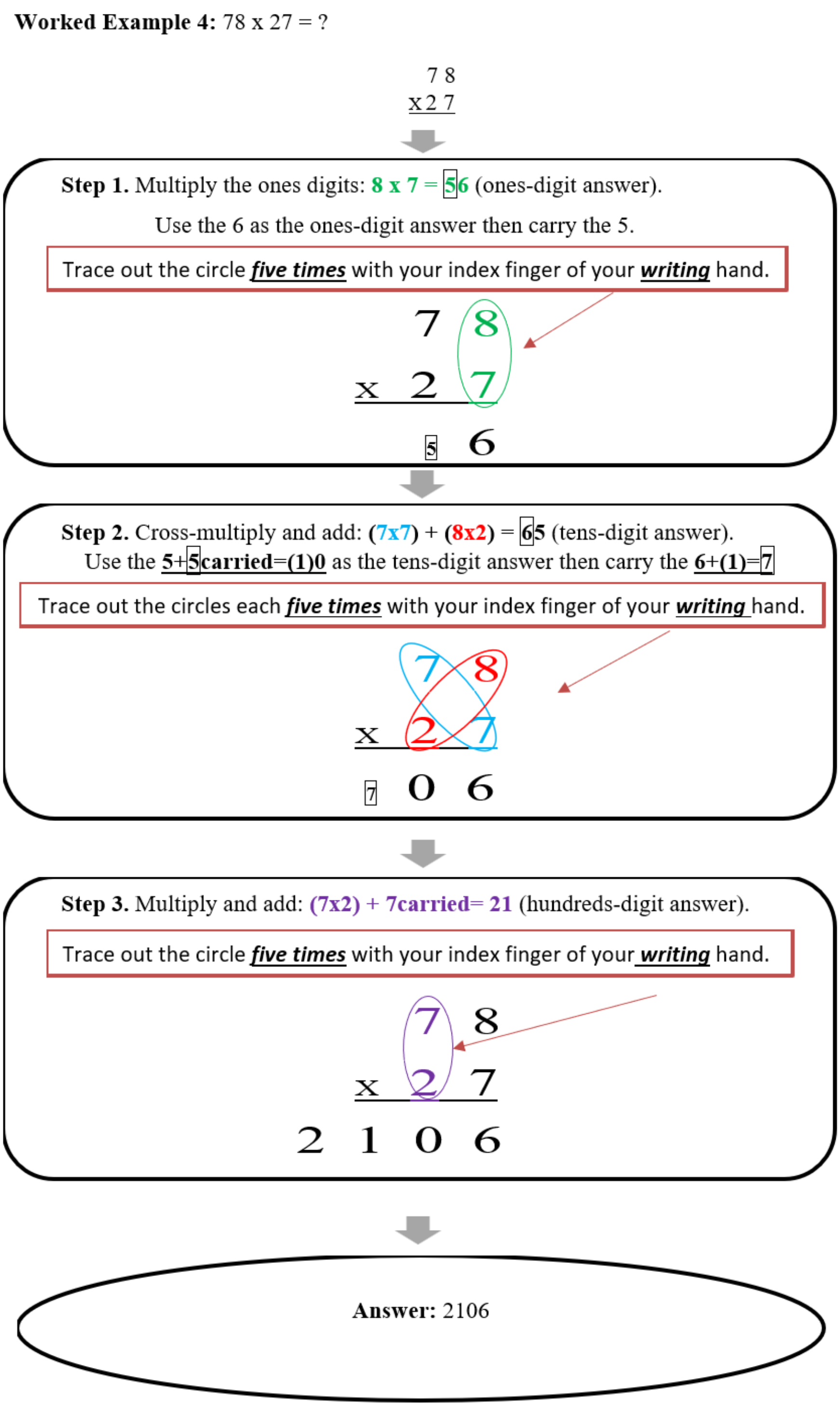


Figure 14

A Worked Example With Two Carry-on Calculations (Complicated) in Experiment 2



Another design detail different to Experiment 1 was the presentation of tracing guidance. In Experiment 1, they were presented in a separate text box that was located on the left side of the page, while in Experiment 2, they were presented within the learning steps that were located in the middle of the page. This adjustment was made because of the potential split attention effect noted in Experiment 1. Figure 15 and Figure 16 show the presenting difference between Experiment 1 and Experiment 2. In each step in Experiment 2, there was a separate text box outlined in red colour that asked participants to trace the key ellipses with the requested index finger.

Figure 15

Examples of Steps and Tracing Guidance in Experiment 1

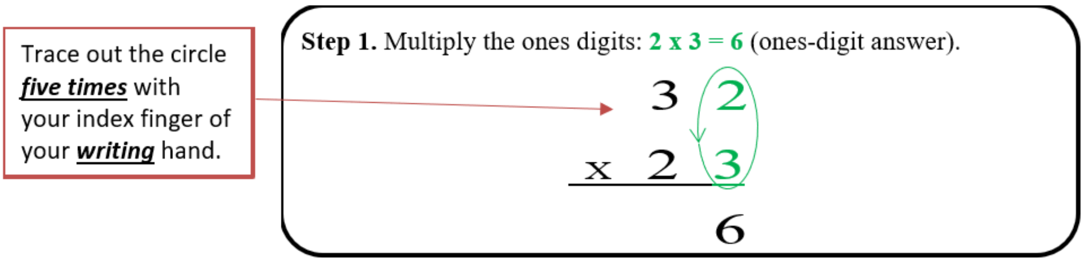
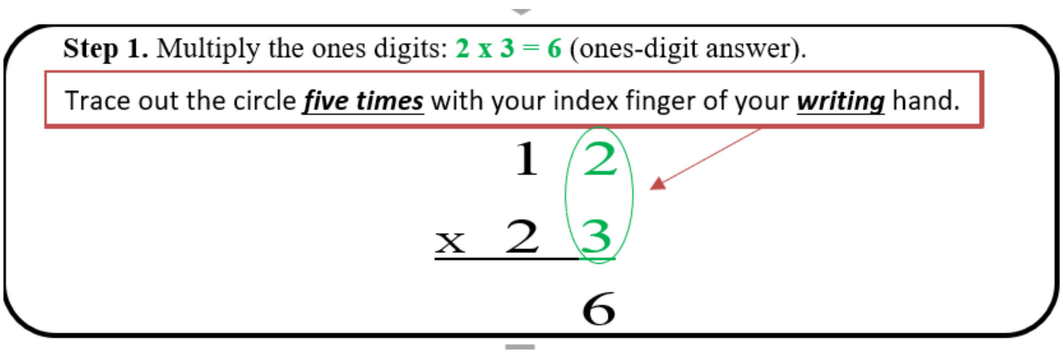


Figure 16

Examples of Steps and Tracing Guidance in Experiment 2



The learning materials were identical for both two conditions except for the separate text box, in which participants were asked to use their writing hand's index finger or the non-writing hand's index finger (according to the conditions) to trace the learning material. Participants in the non-alternating tracing condition were asked to use their writing hand's index finger to trace the learning material all the time, while participants in the alternating tracing condition were asked to switch their writing hand and non-writing hand's index finger to trace the given step.

At the beginning of this phase, participants received a two-page instruction telling them how to learn and trace the learning materials in this phase. The researcher also gave oral instructions to help participants understand the procedure and the requirements in this phase. Participants could take their time to read the instructions until they fully understood the experiment's requirements, and they were encouraged to ask questions for clarification if needed. They were then asked to trace the ellipses on the tracing practice page with guidance from the researcher. Participants then started to learn the two-digit number multiplication method by studying the four worked examples one by one and answering the practice questions that followed each worked example. Participants had two minutes to study each worked example and 30 seconds to answer the following practice question of the corresponding difficulty. The researcher told participants whether their answer was correct or incorrect immediately after participants wrote down the answer. If the answer was wrong, then the participant continued to try until 30 seconds ran out. The sequence was to learn one worked example for two minutes, then answer one practice question, and learn the next worked example for another two minutes, then answer another practice question. During the two-minute study period, participants in both conditions traced the ellipses in the worked example steps with the specified index finger as the separate text box indicated. The tracing instructions in the two conditions were different, specifying the non-alternating index finger (writing hand's index finger) for the non-alternating tracing condition and using alternating index fingers (switch writing and non-writing hands' index fingers) for the alternating tracing condition. During participants' learning time, the researcher sat to one side and watched participants' learning behaviour, reminding them when necessary to focus on learning and use the correct index finger to trace.

Self-report Phase. After the learning phase, participants completed a twenty-item survey on their motivation and different types of cognitive load. Participants needed to rate each statement from 0 (not at all applicable) to 8 (fully applicable), according to their experience in the learning phase. The intrinsic motivation scale was developed by McAuley et al. (1989), including five different items (e.g., "I enjoyed doing this activity very much"). The cognitive load scale (15 items) was developed by Krieglstein et al. (2023), including five items on intrinsic cognitive load (e.g., "The learning content was difficult to understand"), five items on extraneous cognitive load (e.g., " The design of the learning material made it difficult to find relevant information quickly"), and five items on germane processing (e.g., " I actively reflected upon the learning content"). When some participants asked the precise meaning or the referred-to item of some words and phrases in the statement, the researcher read a pre-prepared standardised translation of the statement in response to such questions from the participants.

Compared to Experiment 1, the specific words and sentences of 20 items did not change, but the translation of the statements was more precise and targeted. For example, the statement of item 10 (ICL_5) was "Without prior knowledge, the information was not understandable." In Experiment 1, the translation of this statement was a direct, literal, word-for-word translation. When some participants asked questions like "what did prior knowledge refer to?" and "whether the reading ability to understand the worked example was considered as prior knowledge or not?", the researcher just told participants to follow their own understanding. However, in Experiment 2, the researcher gave them a certain answer that "the prior knowledge referred to math-related knowledge".

Immediate Post-Test Phase. Participants needed to answer 20 two-digit multiplication questions in this phase. For each question, they had 20 seconds to answer. They needed to write down the final answer within the time limit. The researcher did not tell the participants if their answers were correct or not. Participants who finished answering one question just moved to the next one, and the researcher started rerecording the time of the next question. The accuracy of participants' answers and the time they spent solving the questions were recorded to reflect their learning performance. Writing an answer that exceeded 20 seconds was recorded as "out of time", and even if the answer was correct, this question was seen as an unsuccessful answer, which was equal to wrong.

Questions in the post-test phase were a little more difficult than the practice ones in the learning phase, with larger numbers and higher difficulty levels. Practice questions were evenly distributed across the four difficulty levels, with one question for each level as previously described. However, since level 1 was too simple, it was not included in the post-test. Thus, 20 immediate post-test questions were relatively evenly distributed across the remaining levels, with six level 2 questions, nine level 3 questions, and five level 4 questions.

Compared to Experiment 1, which initially only distinguished two levels (worked examples 1-3 with one carry-on calculation and worked example 4 with two carry-on calculations), the current Experiment 2 had more refined difficulty grading, smoother learning progression, and more balanced test question distribution.

Delayed Post-Test Phase. The delayed post-test phase was applied one day (approximately 24 hours) after the initial experiment. Participants needed to answer 20 two-digit multiplication questions again in the delayed post-test phase. The questions used in the delayed post-test were exactly the same as those presented in the immediate post-test.

Debriefing phase. The researcher debriefed the whole experiment to the participants in this phase, clarifying the true purpose of the experiment. Participants were encouraged to express their feelings and thoughts about the experiment and its theoretical assumptions. Some participants asked about their accuracy in the test, and some were interested in the research design. The researcher responded to the participants' questions as much as possible without posing privacy concerns.

7.1.3 Data Processing and Model Check

One hundred and twelve participants joined the experiment, and 111 of them completed the experiment in its entirety. These data constituted the full dataset. As participants' learning processes were fully videotaped, the video recordings can be used to check participants' compliance with instructions. The fully compliant participants constituted the compliant dataset.

To check participants' compliance, the coding scheme used in Experiment 1 was used to quantify participants' effective tracing activities (see Appendix B). Their tracing actions on the required learning materials were recorded as valid counts. One hundred participants were retained after the compliance check, and 12 participants were excluded. Nine participants were removed due to insufficient tracing times (participant number 1, 2, 6, 8, 20, 22, 40, 90, 100), and 3 participants were removed because of incorrect tracing action (participant number 52, 70, 72).

The following data analysis was mostly based on the compliant dataset (100 participants' results). The participant who did not come for the delayed post-test was still in this compliant dataset and was included in most analyses, except those analyses based on delayed post-test data.

The following data analysis was based on the compliant dataset (100 participants' results). The majority of statistical analyses were conducted via jamovi (2.6.26) (The jamovi project, 2024). Detailed data processing procedures were separately presented for each variable.

Handedness Score (Hand Preference and Coordination). As in Experiment 1, participants' hand preference was collected via the handedness survey (Nicholls et al., 2013). All participants needed to report their propensity to use one or both hands. As all participants chose "the right hand" on the first item (With which hand do you write?) and the tenth item (Which hand do you use to draw?), there was no variance on these items. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that the remaining eight items from the handedness survey had low levels of reliability (Cronbach's α = .23; McDonald's ω = .35). Results for this control variable should therefore be viewed with considerable caution.

Pre-Test Score (Prior Calculation Ability). As in Experiment 1, participants' prior abilities were evaluated via their performance in the calculation fluency test. Participants finished a three-page calculation test on addition, subtraction, and multiplication (one test per page). Every correctly answered question on these pages was counted as one point, producing three separate scores for addition, subtraction, and multiplication fluency.

Table 7 presents participants' performance in the calculation fluency test. On average, participants gained a higher score on addition than subtraction, and both of these scores were higher than multiplication. The pre-test scores of participants in Experiment 2 were somewhat lower than those in Experiment 1, but were still consistent with Sowinski et al.'s (2014) results. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that the pre-test scores had suitable levels of reliability (Cronbach's α = .85; McDonald's ω = .92).

As in Experiment 1, though the following learning task in this experiment only involved addition and multiplication in the calculation process, Experiment 2 included subtraction scores together to make the total scores more widely distributed, reflecting participants' calculation ability at a more general level. Scoring the test as the total number of correctly solved items across the three pages was also suggested by the test developer (Sowinski et al., 2014). Thus, three separate scores were added together to generate an overall score to represent their general calculation ability in the following analysis.

Table 7

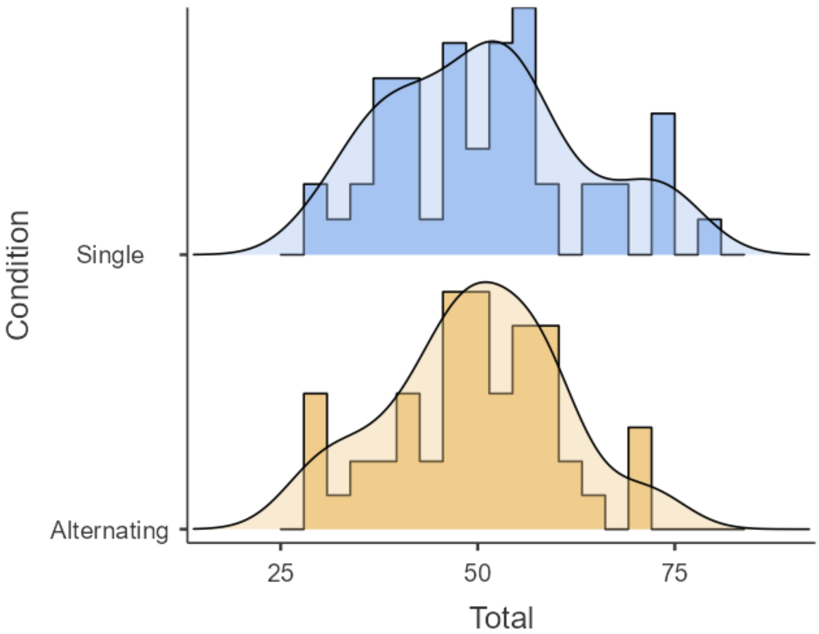
Participants' Performance in the Calculation Fluency Test in Experiment 2

Test Type	n	*M*	*SD*
Addition (/60)	100	20.02	5.12
Subtraction (/60)	100	15.13	4.67
Multiplication (/60)	100	14.90	3.85
Total (/180)	100	50.05	11.68

In this experiment, participants' prior abilities met the assumption of being normally distributed, based on the Shapiro-Wilk (1965) test. This test was considered the most appropriate method for small sample sizes (<50) and was equally effective compared to other methods for medium-sized samples ($50 \leq n < 300$) (Mishra et al., 2019). As this experiment had 100 participants in total and 50 participants in each group, the Shapiro-Wilk test was considered a suitable choice for the normality test. The overall distribution (Shapiro-Wilk $W = .98$, $p = .116$) as well as the distribution of each condition satisfied the normal distribution. As Figure 17 shows, the non-alternating tracing group (Shapiro-Wilk $W = .97$, $p = .148$) had a similar distribution to the alternating tracing group (Shapiro-Wilk $W = .98$, $p = .358$).

Figure 17

The Distribution of Participants' Total Scores in the Pretest in Experiment 2



Practice Phase Learning Performance. As in Experiment 1, participants' learning performance in the learning phase was evaluated via two indicators: the number of practice questions they answered correctly and the time they used to answer the questions. The former was defined as the practice score, and the latter was defined as the practice test time to solution. The practice score was an integer that ranged from 0 to 4. The practice time to solution was obtained by adding up the time participants used to correctly answer each practice question. If the final answer was wrong or participants couldn't give an answer within the time limit (30 seconds), the time of 30 seconds was recorded. Practice time to solution therefore ranged from 0 to 120 seconds.

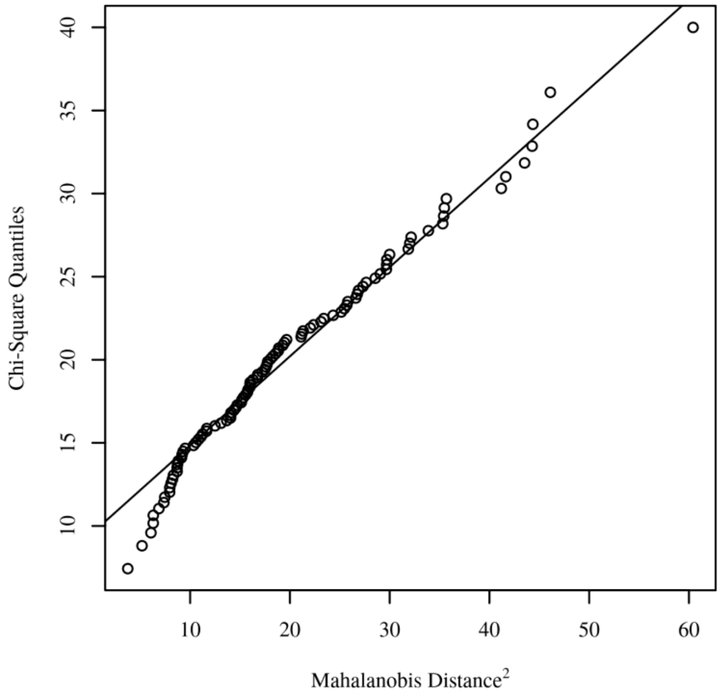
Post-Lesson Self-Report. Participants' self-reports reflected four variables: intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing, each evaluated by five separate statements. The score ranged from 0 to 8, reflecting participants' agreement with the statement.

To check the validity of the current survey results, a confirmatory factor analysis (CFA) was conducted using jamovi (Rosseel et al., 2024). Although CFA results are generally considered to be reliable only with a large sample size (*n* > 300) (Comrey & Lee, 1992), Wolf et al. (2013) proposed the possibility that a CFA model with high factor loadings might be applicable to a sample size lower than 100. Thus, we continued to conduct the CFA model with the current dataset. To check the assumption of multivariate normality, the squared Mahalanobis distances were calculated for the data and plotted against the quantiles of a Chi-square distribution (DeCarlo, 1997). As Figure 18 presents, the assumption of multivariate normality was largely met. Multicollinearity was assessed via the squared multiple correlations and the determinant of the correlation matrix; the item "Motivation3" was found to have an $R^2 > .90$, which indicated the high multicollinearity of this item (Kline, 2015). The Comparative Fit Index (Bentler, 1990), Tucker-Lewis Index (Tucker & Lewis, 1973) and Root Mean Square Error of Approximation (Hu & Bentler, 1999) all showed the model fitted the data well (CFI = .94, TLI = .93, RMSEA = .08).

Compared to Experiment 1, the model fitted better. Further checking the factor loadings (See Table 8), the factor loadings of indicators "ICL_5" and "GCL_2" were too low (< 0.40) to be included in the model (Cabrera-Nguyen, 2010; Matsunaga, 2010). The model was rerun without these indicators, and the results showed a better model fit (CFI = .95, TLI = .94, RMSEA = .08). These indicators were excluded from all the following analyses.

Figure 18

Mahalanobis Distance Scatterplot Testing Multivariate Normality in Experiment 2



Compared to Experiment 1, the indicators "ICL_5" and "GCL_2" loaded on the main factors to a statistically significant level, but they were still excluded from the model because of their low factor loading. Since the statements of the questionnaire did not change, the improvement in the indicator relation was believed to be a result of more precise translation and explanation to the participants.

Table 8

Original Factor Loadings in Experiment 2

Factor	Indicator	Estimate	SE	Z	*p*	Stand. Estimate
Intrinsic Motivation	Motivation1	1.53	0.13	12.00	< .001	0.92
	Motivation2	1.73	0.14	12.59	< .001	0.94
	Motivation3	1.88	0.14	13.01	< .001	0.96
	Motivation4	1.66	0.15	11.18	< .001	0.88
	Motivation5	1.21	0.20	6.17	< .001	0.57
Intrinsic Cognitive Load	ICL_1	1.62	0.16	10.38	< .001	0.85
	ICL_2	1.50	0.14	10.59	< .001	0.86
	ICL_3	1.79	0.14	12.53	< .001	0.94
	ICL_4	1.79	0.15	11.90	< .001	0.92
	ICL_5	0.90	0.23	3.93	< .001	0.39
Extraneous Cognitive Load	ECL_1	1.54	0.15	10.17	< .001	0.84
	ECL_2	1.79	0.17	10.68	< .001	0.87
	ECL_3	1.73	0.18	9.72	< .001	0.82
	ECL_4	1.51	0.20	7.69	< .001	0.69
	ECL_5	1.09	0.20	5.41	< .001	0.52
Germane Processing	GCL_1	1.49	0.17	9.02	< .001	0.79
	GCL_2	0.68	0.23	2.94	.003	0.31
	GCL_3	1.40	0.18	8.00	< .001	0.72
	GCL_4	1.56	0.17	9.15	< .001	0.80
	GCL_5	1.49	0.16	9.23	< .001	0.80

Reliability tests were conducted (Revelle, 2024) to check participants' answers' consistency and dependability. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that four separate factors had suitable levels of reliability (see Table 9). Compared to Experiment 1, all factors had the same or better reliability in Experiment 2.

Participants' motivation and cognitive load were evaluated via the aforementioned 4 factors and 18 items. The average score of each factor was used for the following analyses. All scores ranged from 0 to 8, and higher scores indicate higher motivation or cognitive load.

The Shapiro-Wilk (1965) test was applied to check the normality of participants' average intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing score.

Table 9

Reliability Scores in Experiment 2

Variable	Cronbach's α	McDonald's ω
Intrinsic Motivation (5 items)	.93	.94
Intrinsic Cognitive Load (4 items)	.94	.94
Extraneous Cognitive Load (5 items)	.86	.87
Germane Processing (4 items)	.86	.86

Participants' average intrinsic motivation self-reports (Shapiro-Wilk $W = .95$, $p < .001$) were not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic motivation distribution showed a pronounced left skew: the 25th percentile is 4.2, while the median is as high as 6.0. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic motivation, participant responses of 8, the highest possible response, constituted 17.0%, 26.0%, 21.0%, 13.0%, and 10.0%, for items 1-5 respectively. Based on McHorney and Tarlov's (1995) guidance that ceiling or floor effects are likely to be an issue when more than 15% of respondents respond with either the highest or lowest possible score, a ceiling effect was evident for intrinsic motivation.

Participants' average intrinsic cognitive load self-reports (Shapiro-Wilk $W = .89$, $p < .001$) also were not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic cognitive load distribution showed a pronounced right skew: the median is 1.3 while the 75th percentile is 3.0. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic cognitive load, participant responses of 0, the lowest possible response, constituted 25.0%, 26.0%, 22.0%, and 29.0%, for items 1-4 respectively; thus, a floor effect was evident for self-reports of intrinsic cognitive load (McHorney & Tarlov, 1995).

Participants' average extraneous cognitive load self-reports (Shapiro-Wilk $W = .96$, $p = .008$) also were not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the extraneous cognitive load distribution showed a pronounced right skew: the median is 2.4 while the 75th percentile is 3.6. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For extraneous cognitive load, participant responses of 0, the lowest possible response, constituted 12.0%, 17.0%, 23.0%, 16.0% and 17.0%, for items 1-5 respectively; thus, a floor effect was evident for self-reports of extraneous cognitive load (McHorney & Tarlov, 1995).

Participants' average germane processing self-report (Shapiro-Wilk $W = .92$, $p < .001$) also was not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the motivation distribution showed a pronounced left skew: the 25th percentile is 4.8 while the median is as high as 6. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For germane processing, participant responses of 8, the highest possible response, constituted 15.0%, 16.0%, 24.0%, and 17.0%, for items 1 and 3-5 respectively; thus, a ceiling effect was evident for self-reports of germane processing (McHorney & Tarlov, 1995).

Immediate Post-test Phase Learning Performance. Participant's learning performance in the immediate post-test phase was evaluated via two indicators: the number of test questions they answered correctly, and the time they used to answer the questions. The former was defined as the immediate post-test score, and the latter was defined as the immediate post-test time. The immediate post-test score was an integer that ranged from 0 to 20. The immediate post-test time was obtained by adding the time participants used to correctly answer each test question. If a participant could not give a correct answer within the time limit (20 seconds), then the maximum time of 20 seconds was recorded. The immediate post-test time to solution thus ranged from 0 to 400 seconds.

Delayed Post-test Phase Learning Performance. Participants' learning performance in the delayed post-test phase was evaluated via the same indicators as used in the Immediate Post-test. Thus, the delayed post-test score was an integer that ranged from 0 to 20. As for the Immediate Post-test to solution, the Delayed Post-test time to solution was obtained by adding the time participants used to correctly answer each test question. If a participant could not give a correct answer within the time limit (20 seconds), then the maximum time of 20 seconds was recorded. The delayed Post-test time to solution thus ranged from 0 to 400 seconds.

7.1.4 Statistical Analysis

The procedure and method to test the hypothesis in Experiment 2 were the same as those in Experiment 1. Group equivalence was first checked between the two conditions. ANCOVA and independent-samples *t*-test were used to check the group difference on learning performance, self-reported intrinsic motivation and cognitive load. Tests of statistical significance controlled the Type 1 error rate at 0.05.

7.2 Results

7.2.1 Can We Consider Two Groups of Participants as Equivalent?

In this experiment, participants were randomly assigned to conditions, with the expectation that the participants in two conditions would be equivalent in terms of prior ability.

An independent samples *t*-test was conducted to compare participants' prior calculation abilities. The assumption of homogeneity of variance was satisfied, $F(1, 98)$ =

1.16, $p = .284$. Both the non-alternating group (Shapiro-Wilk $W = 0.97$, $p = .148$) and the alternating group (Shapiro-Wilk $W = 0.98$, $p = .358$) met the assumption of normality. As expected, participants under the two conditions showed no difference in pre-test scores, $t(98) = 0.47$, $p = .640$, $d = 0.11$, 95% CI [-0.30, 0.49], with similar performance in the non-alternating condition ($M = 50.60$, $SD = 12.48$) and the alternating condition ($M = 49.50$, $SD = 10.93$). Thus, participants' prior calculation ability was considered equivalent in the two conditions.

An independent samples *t*-test was conducted to compare participants' handedness preference. The assumption of homogeneity of variance was satisfied, $F(1, 98) = 1.86$, $p = .175$. Though neither the non-alternating group (Shapiro-Wilk $W = 0.74$, $p < .001$) nor the alternating group (Shapiro-Wilk $W = 0.74$, $p < .001$) met the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). Participants' handedness scores were similar across the non-alternating ($M = 0.90$, $SD = 0.14$) and alternating condition ($M = 0.88$, $SD = 0.16$), $t(98) = 0.80$, $p = .429$, $d = 0.16$, 95% CI [-0.23, 0.55].

Fisher's exact test (Upton, 1992) was used to analyse participants' gender distribution across conditions because of its suitability for small sample sizes, instead of the Pearson chi-square test (Pearson, 1900), which required every cell's frequencies to be bigger than five (Kim, 2016). In the current study, there was an option of "prefer not to say" that was selected by only one participant, which did not meet the basic requirements of the chi-squared test. Fisher's exact test indicated that the non-alternating condition and the alternating condition were also equivalent in terms of the distribution of gender, $p = .518$.

7.2.2 Learning Performance

Did Participants' Prior Abilities Affect Their Learning Performance? Though participants' prior calculation ability knowledge can be considered equivalent from the group level, the results of Experiment 1 suggested the possible influence of participants' prior calculation ability on learning from the individual level. To control for this potential influence, we need to confirm again whether individuals' calculation abilities were correlated with their learning performance. A correlational analysis was used to investigate the potential relationship between them. As Table 10 shows, participants' pre-test scores were significantly correlated with all three types of learning performance. Thus, the pre-test scores were considered as a covariate to control for the influence of preexisting differences between individuals in the following analysis.

Table 10

Correlation Between Prior Ability and Learning Performance in Experiment 2

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7
1. Pre-test Score (/180)	100	50.05	11.68	—						
2. Practice Score (/4)	100	2.90	1.07	.30**	—					
3. Practice Time to Solution (/120)	100	91.29	16.97	-.40***	-.78***	—				
4. Immediate Post-Test Score (/20)	100	11.84	4.63	.28**	.36***	-.48***	—			
5. Immediate Post-Test Time to Solution (/400)	100	347.19	32.67	-.45***	-.40***	.56***	-.82***	—		
6. Delayed Post-Test Score (/20)	99	16.23	4.09	.24*	.35***	-.41***	.68***	-.50***	—	
7. Delayed Post-Test Time to Solution (/400)	99	301.54	39.96	-.42***	-.32**	.51***	-.69***	.78***	-.76***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

How Did Participants Perform in the Practice Phase? Since participants' prior abilities were found to correlate with their learning performance in the practice phase, the ANCOVA test (Fox et al., 2024) was used to compare participants' learning performance in the practice phase across the two conditions. Statistical analyses controlled the Type 1 error rate at 0.05. Two indicators, practice score and practice time, were separately analysed with the same data processing procedure.

For the practice score, homogeneity of variances was found via the Levene (1960) test, $F(1, 98) = 0.25$, $p = .618$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.11$, $p = .003$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, referring to Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. The interaction effect between condition and prior abilities was not statistically significant, $F(1, 96) = 2.54$, $p = .114$, $\eta^2p = .03$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. As expected, the main effect of prior calculation ability was statistically significant, $F(1, 97) = 9.73$, $p = .002$, $\eta^2p = .09$; participants with higher pre-test scores also gained higher practice scores. The main effect of the condition was not statistically significant, $F(1, 97) = 0.06$, $p = .810$, $\eta^2p = .00$; participants in the non-alternating condition (estimated marginal $M = 2.93$, $SE = .15$) solved similar problems correctly at a similar rate to those in the alternating condition (estimated marginal $M = 2.88$, $SE = .15$).

Since the mean practice scores were high in both conditions, potential ceiling effects may exist in the practice scores. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For the practice score, participant score of 4, the highest possible score, constituted 42.0% and 28.0% for non-alternating and alternating conditions respectively. Based on McHorney and Tarlov's (1995) guidance that ceiling or floor effects are likely to be an issue when more than 15% of respondents respond with either the highest or lowest possible score, a ceiling effect was evident for the practice score. Meier and Feeley's (2022) criterion, based on the standard deviation, also suggested the presence of a ceiling effect as the sum of the mean plus 1-2 *SD* equalled or exceeded the maximum test score. For the practice score, the sum of the mean plus 1 *SD* was 3.97, and the sum of the mean plus 2 *SD* was 5.04, further confirming the existence of the ceiling effect.

Considering the ceiling effect was detected on the delayed post-test score, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the result against this potential assumption violation. Under this analysis, $t(98) = 0.43$, $p = .673$, $d = 0.11$, 95% CI [-0.50, 0.76], indicating that there was indeed no significant difference in participants' delayed post-test scores between the non-alternating finger tracing group (corrected $M = 3.07$, $SD = 1.23$) and the alternating fingers tracing group (corrected $M = 2.94$, $SD = 1.10$).

For practice time to solution, homogeneity of variances was found via Levene's (1960) test, $F(1, 98) = 2.50$, $p = .117$, and normality of residuals was found via the Lilliefors (1967) test ($D = 0.06$, $p = .438$). The interaction effect between the condition and the prior abilities was not statistically significant, $F(1, 96) = 0.16$, $p = .692$, $\eta^2 p = .00$, showing that the assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 97) = 18.26$, $p < .001$, $\eta^2p = .16$; participants with higher pre-test scores spent less time finishing practice questions. The main effect of condition was not statistically significant, $F(1, 97) = 0.149$, $p = .700$, $\eta^2p = .00$; participants in the non-alternating condition (estimated marginal $M = 90.68$, $SE = 2.22$) spent similar time solving problems to those in the alternating condition (estimated marginal $M = 91.90$, $SE = 2.22$).

Contrary to Hypothesis *H1*, both indicators showed that participants in the non-alternating condition performed approximately the same as those in the alternating condition on the practice questions.

How Did Participants Perform in the Immediate Post-Test Phase? The ANCOVA test (Fox et al., 2024) was used to compare participants' learning performance in the immediate post-test in the two conditions. Analyses controlled the Type 1 error rate at 0.05. Two indicators, immediate post-test score and immediate post-test time to solution, were separately analysed with the same data processing procedure.

For the immediate post-test score, homogeneity of variances was found via the Levene (1960) test, $F(1, 98) = 0.90$, $p = .344$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.14$, $p < .001$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. No interaction effect was found between the condition and the prior abilities on the immediate post-test scores, $F(1, 96) = 1.69$, $p = .196$, $\eta^2p = .02$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for immediate post-test scores, $F(1, 97) = 7.75$, $p = .006$, $\eta^2p = .07$; participants with higher pre-test scores also gained higher immediate post-test scores. The main effect of the condition was not significant for immediate post-test scores, $F(1, 97) = 0.80$, $p = .374$, $\eta^2p = .01$; participants in the non-alternating condition (estimated marginal $M = 12.24$, $SE = 0.63$) gained similar scores to those in the alternating condition (estimated marginal $M = 11.44$, $SE = 0.63$).

For immediate post-test time to solution, homogeneity of variances was found via the Levene (1960) test, $F(1, 98) = 1.60$, $p = .209$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.05$, $p = .698$). No interaction effect was found between the condition and the prior abilities on the immediate post-test time to solution, $F(1, 96) = 0.56$, $p = .455$, $\eta^2p = .01$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for immediate post-test time to solution, $F(1, 97) = 24.85$, $p < .001$, $\eta^2p = .20$, showing that participants with higher pre-test scores spent less time finishing immediate post-test questions. The main effect of the condition on immediate post-test time to solution was also not significant, $F(1, 97) = 0.01$, $p = .940$, $\eta^2p = .00$; participants in the non-alternating condition (estimated marginal $M = 346.97$, $SE = 4.17$) spent similar time solving problems to those in the alternating condition (estimated marginal $M = 347.41$, $SE = 4.17$).

Contrary to Hypothesis *H6*, both indicators showed that participants in the non-alternating condition performed similarly to those in the alternating condition on the immediate post-test questions.

How Did Participants Perform in the Delayed Post-Test Phase? Procedures to analyse participants' learning performance in the delayed post-test phase were similar to those used for the immediate post-test.

For the delayed post-test score, homogeneity of variances was found via the Levene (1960) test, $F(1, 97) = 0.43$, $p = .512$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.18$, $p < .001$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. No interaction effect was found between the condition and the prior abilities on the delayed post-test scores, $F(1, 95) = 0.96$, $p = .330$, $\eta^2p = .01$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for delayed post-test scores, $F(1, 96) = 5.75$, $p = .018$, $\eta^2p = .06$; participants with higher pre-test scores also gained higher delayed post-test scores. The main effect of the condition was not significant for delayed post-test scores, $F(1, 96) = 0.38$, $p = .539$, $\eta^2p = .00$; participants in the non-alternating condition gained similar scores (estimated marginal M = 16.48, SE = 0.57) to those in the alternating condition (estimated marginal M = 15.98, SE = 0.57).

Since the mean delayed post-test scores were high in both conditions, potential ceiling effects may exist in the delayed post-test scores. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For the delayed post-test score, participant score of 20, the highest possible score, constituted 12.0%

and 10.2% for non-alternating and alternating conditions respectively. If the criteria were relaxed to the highest two scores, participant scores of 19 or 20 constituted 30.0% and 24.5% for the non-alternating and alternating conditions respectively. Based on McHorney and Tarlov's (1995) guidance that ceiling or floor effects are likely to be an issue when more than 15% of respondents respond with either the highest or lowest possible score, a ceiling effect was evident for the delayed post-test score. Meier and Feeley's (2022) criterion, based on the standard deviation, also suggested the presence of a ceiling effect as the sum of the mean plus 1-2 *SD* equalled or exceeded the maximum test score. For the delayed post-test score, the sum of the mean plus 1 *SD* was 20.32, further confirming the existence of the ceiling effect.

Considering the ceiling effect was detected on the delayed post-test score, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the result against this potential assumption violation. Under this analysis, $t(98) = 0.69$, $p = .492$, $d = 0.15$, 95% CI [-1.10, 2.24], indicating that there was indeed no significant difference in participants' delayed post-test scores between the non-alternating finger tracing group (corrected $M = 17.15$, $SD = 4.16$) and the alternating fingers tracing group (corrected $M = 16.58$, $SD = 3.46$).

For the delayed post-test time to solution, homogeneity of variances was found via the Levene (1960) test, $F(1, 97) = 2.23$, $p = .139$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.06$, $p = .412$). No interaction effect was found between the condition and the prior abilities on the delayed post-test time to solution, $F(1, 95) = 0.20$, $p = .652$, $\eta^2p = .00$; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was significant for delayed post-test time to solution, $F(1, 96) = 20.50$, $p < .001$, $\eta^2 p = .18$, showing that participants with higher pre-test scores spent less time finishing delayed post-test questions. The main effect of the condition on delayed post-test time to solution was also not significant, $F(1, 96) = 0.02$, $p = .894$, $\eta^2 p = .00$; the participants in the non-alternating condition spent similar time solving problems (estimated marginal $M = 301.05$, $SE = 5.18$) to those in the alternating condition (estimated marginal $M = 302.03$, $SE = 5.24$).

Contrary to Hypothesis *H7*, both indicators showed that participants in the non-alternating condition performed similarly to those in the alternating condition on the delayed post-test questions.

How Did Participants' Learning Performance Change Between the Immediate Post-Test and the Delayed Post-Test? To further analyse the change of participants' mastery level between the immediate post-test and the delayed post-test, a repeated measures ANCOVA (Algina, 1982) was conducted. The pre-test score was used as a covariate because it was correlated with participants' immediate post-test performance and delayed post-test performance. Test time was the within-subject factor in the analysis and the key factor this analysis focused on. Two indicators, the test score and the time to solution, were separately analysed with the same data processing procedure.

For the test score, homogeneity of variances was found via the Levene (1960) test both in the immediate post-test, $F(1, 97) = 0.76$, $p = .387$, and the delayed post-test, $F(1, 97) = 0.43$, $p = .512$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.16$, $p < .001$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on

Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

The main effect for test time (immediate vs. delayed post-test) was significant, $F(1, 96) = 159.32$, $p < .001$, $\eta^2 p = .62$, indicating there were significant differences between the immediate post-test score and delayed post-test score; participants gained higher test scores in the delayed post-test (estimated marginal $M = 16.23$, $SE = 0.40$) than in the immediate post-test (estimated marginal $M = 11.75$, $SE = 0.44$). The main effect for the condition was not significant, $F(1, 96) = 0.92$, $p = .339$, $\eta^2 p = .01$, indicating participants in the non-alternating condition gained similar test scores (estimated marginal $M = 14.36$, $SE = 0.54$) to those in the alternating condition (estimated marginal $M = 13.62$, $SE = 0.55$). The main effect for pre-test score was significant, $F(1, 96) = 8.54$, $p < .001$, $\eta^2 p = .08$, indicating participants with higher prior calculation ability also gained higher test scores. The interaction effect between the test time and condition was not significant, $F(1, 96) = 0.47$, $p = .495$, $\eta^2 p = .01$, indicating that participants allocated to different conditions had similar trends of change on their test scores between the immediate and delayed post-tests. The interaction effect between the test time and pre-test scores was also not significant, $F(1, 96) = 0.80$, $p = .375$, $\eta^2 p = .01$, indicating that participants with different prior calculation ability have similar trends of change between the immediate and delayed post-tests.

For the time to solution, homogeneity of variances was found via the Levene (1960) test both in the immediate post-test, $F(1, 97) = 1.26$, $p = .265$, and the delayed post-test, $F(1, 97) = 2.23$, $p = .139$, meeting the assumption of homoscedasticity. The normality of residuals was also found via the Lilliefors (1967) test ($D = 0.04$, $p = .459$).

The main effect for the test time was significant, $F(1, 96) = 331.76$, $p < .001$, $\eta^2 p = .78$, indicating there were significant differences between the immediate post-test time to solution and delayed post-test time to solution; participants spent less time to completing the test questions in the delayed post-test (estimated marginal $M = 301.54$, $SE = 3.68$) than in the immediate post-test (estimated marginal $M = 347.59$, $SE = 2.94$). The main effect for the condition was not significant, $F(1, 96) = 0.04$, $p = .851$, $\eta^2 p = .00$, indicating participants in the non-alternating condition spent similar time completing the test questions (estimated marginal $M = 323.99$, $SE = 4.34$) than those in the alternating condition (estimated marginal $M = 325.15$, $SE = 4.38$). The main effect for pre-test score was significant, $F(1, 92) = 26.20$, $p < .001$, $\eta^2 p = .21$, indicating participants with higher prior calculation ability spent less time completing the test questions. The interaction effect between the test time and condition was not significant, $F(1, 96) = 0.01$, $p = .944$, $\eta^2 p = .00$, indicating that participants allocated to different conditions had similar trends of change on their time to solution between the immediate and delayed post-tests. The interaction effect between the test time and pre-test scores was also not significant, $F(1, 96) = 0.50$, $p = .480$, $\eta^2 p = .01$, indicating that participants with different prior calculation ability have similar trends of change in their time to solution between the immediate and delayed post-tests.

7.2.3 Post-Lesson Self-Report

Did Participants' Prior Abilities Affect Their Motivation and Cognitive Load? As for learning performance results reported above, a correlational analysis was also conducted to determine prior ability's relations with motivation and cognitive load. Table 11 presents detailed results of these analyses. Neither intrinsic motivation nor any kind of cognitive load was found to be significantly correlated with pre-test scores. Thus, an independent *t*-test was applied to check the group difference of participants' first-stage intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing.

Table 11

Correlation Between Prior Ability, Motivation, and Cognitive Load in Experiment 2

Variable	n	M	SD	1	2	3	4	5
1. Pre-Test Score (/180)	100	50.23	11.63	—				
2. Intrinsic Motivation (/8)	100	5.58	1.68	.08	—			
3. Intrinsic Cognitive Load (/8)	100	1.88	1.73	-.12	-.50***	—		
4. Extraneous Cognitive Load (/8)	100	2.56	1.66	-.07	-.49***	.78***	—	
5. Germane Processing (/8)	100	5.74	1.61	-.09	.63***	-.66***	-.65***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

What is the Intrinsic Motivation Reported by Participants? Since correlational analysis showed there was no clear correlation between participants' prior knowledge and their intrinsic motivation, an independent-samples *t*-test was applied to analyse participants' self-reported intrinsic motivation.

The assumption of homogeneity of variance was satisfied, $F(1, 98) = 0.03$, $p = .867$. Though neither the non-alternating group (Shapiro-Wilk $W = 0.94$, $p = .016$) nor the alternating group (Shapiro-Wilk $W = 0.93$, $p = .007$) met the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). The difference in self-reported intrinsic motivation ratings reported by the non-alternating condition ($M = 5.73$, $SD = 1.69$) and the alternating condition ($M = 5.43$, $SD = 1.67$) was not statistically significant, $t(98) = 0.89$, $p = .374$, $d = 0.18$, 95% CI [-0.22, 0.57].

Considering the ceiling effect was detected on the average intrinsic motivation scores, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the result against this potential assumption violation. Under this analysis, $t(98) = 1.02$, $p = .311$, $d = 0.21$, 95% CI [-0.37, 1.14], indicating that there was indeed no significant difference in participants' self-reported intrinsic motivation between the non-alternating finger tracing group (corrected $M = 5.85$, $SD = 1.88$) and the alternating fingers tracing group (corrected $M = 5.46$, $SD = 1.74$).

Overall, participants in both groups reported relatively high levels of intrinsic motivation. However, contrary to *H2*, there was no significant difference in their motivation during the lesson between groups.

What is the Intrinsic Cognitive Load Reported by Participants? Since correlational analysis showed there was no clear correlation between participants' prior calculation knowledge and their intrinsic cognitive load, an independent-samples *t*-test was applied to investigate group differences in participants' self-reported intrinsic cognitive load.

The assumption of homogeneity of variance was satisfied, $F(1, 98) = 1.13$, $p = .291$. Though neither the non-alternating group (Shapiro-Wilk $W = 0.86$, $p < .001$) nor the alternating group (Shapiro-Wilk $W = 0.91$, $p < .001$) met the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). The difference in self-reported intrinsic cognitive load ratings reported by the non-alternating condition ($M = 1.64$, $SD = 1.61$) and the alternating condition ($M = 2.11$, $SD = 1.83$) was not statistically significant, $t(98) = -1.36$, $p = .177$, $d = -0.27$, 95% CI [-0.67, 0.12].

Considering the floor effect was detected on the average intrinsic cognitive load scores, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the result against this potential assumption violation. Under this analysis, $t(98) = -2.13$, $p = .040$, $d = -0.47$, 95% CI [-0.98, -0.03], indicating participants in the alternating group (corrected $M = 2.18$, $SD = 1.08$) perceived higher intrinsic cognitive load than those in the non-alternating group (corrected $M = 1.67$, $SD = 1.08$).

Overall, participants in both groups reported relatively low levels of intrinsic cognitive load, with a floor effect apparent in participants' reports of intrinsic cognitive load. Aligning with *H3,* the average intrinsic cognitive load between the non-alternating tracing group and the alternating tracing group showed a significant difference after considering the floor effect, suggesting participants in the non-alternating finger tracing group perceived lower intrinsic cognitive load.

What is the Extraneous Cognitive Load Reported by Participants? Since correlational analysis showed there was no correlation between participants' prior knowledge and their extraneous cognitive load, an independent-samples *t*-test was applied to investigate group differences in participants' extraneous cognitive load.

The assumption of normality was met in the alternating group (Shapiro-Wilk W = 0.96, p = .113), but not in the non-alternating group (Shapiro-Wilk W = 0.95, p = .042). The assumption of homogeneity of variance was satisfied, $F(1, 98) = 0.01$, p = .912. Though one of the conditions did not meet the assumption of normality, the *t*-test was still considered to be valid with a larger sample size (n > 30) (Box, 1953; Lumley et al., 2002). The difference in self-reported extraneous cognitive load ratings reported by the alternating condition (M = 1.45, SD = 1.35) and the non-alternating condition (M = 1.90, SD = 1.37) was statistically significant, $t(98) = -2.05$, p = .043, d = -0.41, 95% CI [-0.80, -0.01]. Given the apparent floor effect for this variate, Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects was used to check the result against this potential assumption violation. Under this analysis, $t(98) = -4.20$, $p < .001$, d = -0.86, 95% CI [-0.97, -0.34], indicating participants in the alternating group (corrected M = 2.95, SD = 0.73) perceived higher extraneous cognitive load than those in the non-alternating group (corrected M = 2.29, SD = 0.79).

Overall, participants in both groups reported relatively low levels of extraneous cognitive load, with a floor effect apparent in participants' reports of extraneous cognitive load. Contrary to *H4,* the average extraneous cognitive load between the non-alternating finger tracing group and the alternating-finger tracing group showed a significant difference after considering the floor effect, suggesting participants in the non-alternating finger tracing group perceived lower extraneous cognitive load.

What is the Germane Processing Reported by Participants? Since correlational analysis showed there was no clear correlation between participants' prior calculation ability and their germane processing scores, an independent-samples *t*-test was applied to investigate group differences in participants' reports of germane processing.

The assumption of homogeneity of variance was satisfied, $F(1, 98) = 1.94$, $p = .166$. Though neither the non-alternating group (Shapiro-Wilk $W = 0.93$, $p = .004$) nor the alternating group (Shapiro-Wilk $W = 0.92$, $p = .003$) met the assumption of normality, the *t*-test was still considered to be valid with a larger sample size ($n > 30$) (Box, 1953; Lumley et al., 2002). The difference in self-reported germane processing ratings reported by the non-alternating condition ($M = 5.95$, $SD = 1.44$) and the alternating condition ($M = 5.53$, $SD = 1.75$) was not statistically significant, $t(98) = 1.33$, $p = .188$, $d = 0.27$, 95% CI [-0.13, 0.66].

Considering the ceiling effect was detected on the average germane processing scores, we used Liu and Wang's (2021) *t*-test correcting for ceiling effects and floor effects to check the data to check the result against this potential assumption violation. Under this analysis, $t(98) = 1.12$, $p = .268$, $d = 0.23$, 95% CI [-0.32, 1.11], indicating that there was indeed no significant difference in participants' self-reported germane processing between the non-alternating finger tracing (corrected $M = 6.00$, $SD = 1.54$) group and the alternating fingers tracing group (corrected $M = 5.61$, $SD = 1.90$).

Overall, participants in both groups reported relatively high levels of germane processing. However, contrary to *H5*, there was no significant difference in their germane processing scores between groups.

7.3 Discussion

7.3.1 How Did Tracing Conditions Affect Participants' Learning Performance?

The above analyses of participants' learning performance in the practice phase, immediate post-test phase, and delayed post-test phase presented similar results, indicating that the alternating finger tracing strategy resulted in similar levels of performance as the non-alternating finger tracing strategy. Alternating tracing actions were designed to further support participants' learning, but the current results did not support this hypothesis. To investigate why the results did not align with expectations, we further summarised the current findings.

Overall, participants' accuracy rates (see Table 12) exhibited a "✓" shaped curve across the three phases. Participants in both groups had moderate accuracy in the practice phase, lower accuracy in the immediate post-test phase, and highest accuracy in the delayed post-test phase. Repeated-measure analysis also suggested that participants' performance in the delayed post-test was significantly better than in the immediate post-test. These findings were aligned with the trend we found in Experiment 1. Better performance in the delayed post-test suggested the possible presence of a practice effect (Anastasi, 1988; Campbell & Stanley, 1963), especially considering that the participants completed identical questions in the immediate post-test and the delayed post-test.

Table 12

Participants' Accuracy Rates in Three Tests in Experiment 2

Performance	Total	Non-Alternating	Alternating
Practice Questions (4 points)			
Mean Score	2.89	2.96	2.82
Accuracy Rate	72.25%	73.98%	70.60%
Immediate Post-Test (20 points)			
Mean Score	11.70	12.27	11.16
Accuracy Rate	58.50%	61.33%	55.79%
Delayed Post-Test (20 points)			
Mean Score	16.17	16.53	15.82
Accuracy Rate	80.86%	82.66%	79.10%

In group comparisons, participants in the non-alternating condition generally had similar performance to the participants in the alternating condition in practice questions, immediate post-test questions, and delayed post-test questions. Considering the aforementioned trends, we can infer that the non-alternating finger tracing strategy and the alternating fingers tracing strategy had similar effects on promoting learning. Participants may have forgotten some problem-solving methods when they first encountered immediate post-test questions, which explained why their accuracy rate went down at this stage. As participants completed the problems while practising, they recalled and better mastered the problem-solving strategies, which resulted in their increasing accuracy rate.

In contrast to Experiment 1's finding that the superior performance of the non-alternating condition gradually diminished across three phases, Experiment 2 did not find any clear changes in the between-condition differences across the three question phases. The non-alternating finger tracing strategy and the alternating fingers tracing strategy seemed to have similar supportive effects and effect duration time in Experiment 2. The two possible explanations proposed in Experiment 1 may also explain the results of Experiment 2. It is possible that the effects of the tracing interventions were no longer effective after 24 hours, and non-alternating finger tracing had similar supportive effects as alternating fingers tracing, which resulted in participants' similar performance in all phases. It is also possible that the effect of alternating tracing will take longer to manifest, meaning we cannot draw firm conclusions because of a 24-hour delay design.

7.3.2 How Did Tracing Conditions Affect Participants' Intrinsic Motivation and Cognitive Load?

Previous analysis described participants' intrinsic motivation and three types of cognitive load in the non-alternating tracing and the alternating tracing conditions. Participants in the non-alternating group reported similar intrinsic motivation and similar germane processing to those in the alternating group. Significant difference was found on intrinsic cognitive load and extraneous cognitive load: participants in the alternating group perceived higher intrinsic and extraneous cognitive load than those in the non-alternating group. These findings were partly different from the findings in Experiment 1 and partly aligned with the current research's initial hypothesis. Table 13 presents research hypotheses and current findings in two experiments.

In contrast to the interaction effect found in Experiment 1, a significant difference was found in participants' intrinsic cognitive load between the two conditions in Experiment 2. As previously mentioned, Experiment 2 adjusted the tracing practice page to help participants better understand the requirements and modified specific worked examples to make the learning process more gradual. These adjustments could make participants feel that alternating finger tracing was less difficult, especially for those participants with low calculation ability. Though participants in the alternating group still perceived higher intrinsic cognitive load, the difference between participants with different prior calculation abilities was no longer significant, resulting in the disappearance of the interaction effect. In Experiment 1, when participants in the alternating group were asked to switch their index fingers during the worked example, some participants may have been overwhelmed and unclear as to whether they should put the learning emphasis on the two-digit multiplication method or the pattern of finger switching. In contrast, in Experiment 2, they practised finger switching on the tracing practice page; thus, they could better understand the key points of learning. For participants in the non-alternating group in both experiments, their learning process never involved complex finger-switching activities, and thus they could better know that the learning focus was the two-digit multiplication. Thus, participants in the alternating group in Experiment 2 generally reported higher intrinsic cognitive load than the non-alternating group, in contrast to Experiment 1.

Table 13

Research Hypotheses on Motivation and Cognitive Load and Corresponding Results in Experiment 1 and Experiment 2

Variable	Hypothesis	Experiment 1	Experiment 2
Intrinsic Motivation	Alternating Higher	Same	Same
Intrinsic Cognitive Load	Alternating Higher	Interaction Effect	Alternating Higher
Extraneous Cognitive Load	Same	Alternating Higher	Alternating Higher
Germane Processing	Alternating Higher	Same	Same

As in Experiment 1, participants in the alternating fingers tracing group reported higher extraneous cognitive load, which was unexpected. The initial hypothesis posited that extraneous load should be similar between the two conditions. Experiment 2 adjusted the presentation of tracing guidance to avoid the potential split attention effect (Chandler & Sweller, 1991) that we suspected may have existed in Experiment 1. Experiment 2 referenced Wang's (2022) design and placed these instructions right below the learning steps in the middle of the page, with the goal of helping participants avoid unnecessary search and match eye movements. We tried to adjust the presentation to avoid the split attention effect; however, the reported extraneous cognitive load of participants in the alternating group was still significantly higher than those in the non-alternating group. We also found that participants' average extraneous cognitive load in both groups (*M* non-alternating = 2.22, *M* alternating = 2.89) in Experiment 2 were higher than that in Experiment 1 (*M* non-alternating = 1.45, *M alternating* = 1.90), indicating that adjusted tracing guidance may have caused higher extraneous cognitive load. Although the participants in Experiment 1 and Experiment 2 were completely different groups of people, the extraneous cognitive load should technically only be related to the learning materials' presentation (Sweller, van Merrienboer, Paas, 1998; Sweller, Ayres, & Kalyuga, 2011). These comparisons between Experiment 1 and Experiment 2 suggested another possibility: that stacking the tracing guidance with the learning steps together in the middle may actually not reduce the cognitive load but rather increase it. Although participants in Experiment 2 no longer needed to move their gaze left and right to read tracing guidance, the split-attention effect may still exist, as they still needed to move their gaze up and down to read the learning steps and the tracing guidance. To better design the learning and reduce unnecessary extraneous cognitive load, the presentation of the tracing guidance may need further adjustment in the following experiment. Possible modifications include simplifying the tracing guidance so that participants can focus more on the explanation of the learning steps during learning and are no longer distracted by additional information.

As in Experiment 1, no difference in germane processing was detected in Experiment 2. We initially expected that participants in the alternating fingers tracing condition might report higher germane processing, because the alternating action would help them clearly distinguish the learning steps and facilitate information processing. However, the only thing participants perceived was the complexity of the alternating actions, not their underlying hints for learning. Due to the alternating fingers tracing design not achieving its intended effect, participants in the non-alternating group and alternating group perceived similar levels of germane processing.

Summarising results against hypotheses across both Experiment 1 and Experiment 2, hypotheses have generally not been supported. The only finding that supported the hypothesis is that participants in the alternating condition reported higher intrinsic cognitive load than those in the non-alternating condition in Experiment 2. Alternating tracing actions across the hands within a specific worked example was hypothesised to result in higher intrinsic cognitive load and higher germane processing, and subsequently, better test performance. However, Experiment 1 and Experiment 2 demonstrated that simply having a more variable tracing process did not directly lead to higher germane processing, reflected by the fact that the interaction effect was detected for intrinsic cognitive load but not on germane processing in Experiment 1, and higher intrinsic cognitive load in the alternating condition but similar germane processing in two conditions in Experiment 2. A similar finding was also suggested by Klepsch and Seufert (2020), that higher intrinsic cognitive load does not necessarily result in higher germane processing. Focusing on the current research, if participants did not understand the alternating fingers tracings' underlying hints on learning and only viewed the alternating action as complexity or interference, they would fail to benefit from it. Furthermore, they may alternatively experience higher extraneous cognitive load. As Sweller (2010) proposed, the same information can impose intrinsic or extraneous cognitive load depending on what needs to be learned. In the current research, if participants do not recognise that the action of "alternating fingers" itself and the underlying hints of it were part of the learning content, and only treat the action as an unnecessary burden, then it is reasonable that they would report higher extraneous cognitive load in the self-estimated survey.

Overall, the results of Experiments 1 and 2 suggest that alternative designs to incorporate variability in tracing actions should be considered. Accordingly, Experiment 3 was designed to investigate a novel approach to alternating finger tracing and its effects on learning processes and outcomes.

Chapter Eight: Experiment 3

As the results of Experiment 1 and Experiment 2 suggested potential for improvement with the design of lesson materials, another experiment was planned to further investigate the effect of alternating-fingers tracing on learning and cognitive load. A number of modifications were formulated for the experimental procedure. Specifically, the presentation of the tracing introduction, the tracing strategy in each condition, the sequence and numbers of the worked examples, and the difficulty level of the immediate and delayed post-tests were adjusted.

Experiment 3 divided the acquisition phase into two stages. In the first stage, participants in both conditions applied the non-alternating tracing strategy. In the second stage, participants applied the non-alternating or alternating tracing strategy according to their allocated condition. This design was inspired by Wang et.al. ’s (2022) two-stage sequence, and aimed to provide participants a solid foundation and clear understanding of the worked example-based mental mathematics strategy in the first stage. Participants applied non-alternating tracing in the first stage, becoming familiar with the tracing action and the learning material. The experimental intervention started in the second stage, with participants applying different tracing strategies (alternating vs. non-alternating). We expected that, having partially developed a schema for the mental mathematics strategy in the first phase, students in the alternating condition would further develop and automate the mental mathematics strategy in the second stage to a greater extent than participants in the non-alternating condition, because of the higher (but manageable) intrinsic load and germane processing generated by alternating tracing. The adjustments in the acquisition phase also impacted the self-report phase. Participants completed the 20-item survey on their intrinsic motivation and cognitive load twice in Experiment 3, following each learning stage. Experiment 3 also adjusted the test questions in the immediate post-test and delayed post-test phase, classifying the questions into similar questions and transfer questions. More detailed adjustments are described in the following sections.

Redesign of Experiment 3 also influenced the analyses of the various self-reports. As all participants applied the same non-alternating tracing strategy in the first stage and they were assumed to be similar across conditions, the repeated measure 2 (non-alternating vs. alternating) x 2 (first stage vs. second stage) AN(C)OVA analyses were more suitable for Experiment 3.

These adjustments have implications for hypotheses and analyses of learning performance and self-reports. The hypotheses below were adjusted for Experiment 3:

H1: Participants' problem-solving performance in the first-stage acquisition phase will not differ across conditions (*H1a*). Participants' problem-solving performance in the second-stage acquisition phase will differ across conditions, reflected by problem-solving accuracy and problem-solving time. Specifically, participants in the non-alternating group will solve fewer problems and/or solve problems more slowly than participants in the alternating group (*H1b*).

H2: The interaction effect between condition and learning stage will be significant on participants' self-reported intrinsic motivation. Specifically, participants in the non-alternating group will report lower second-stage intrinsic motivation than participants in the alternating group (*H2a*).

H3: The interaction effect between condition and learning stage will be significant on participants' self-reported intrinsic cognitive load. Specifically, participants in the non-alternating group will report lower second-stage intrinsic cognitive load than participants in the alternating group (*H3a*).

H4: The interaction effect between condition and learning stage will not be significant on participants' self-reported extraneous cognitive load. Participants' second-stage self-reported extraneous cognitive load will not differ across conditions (*H4a*).

H5: The interaction effect between condition and learning stage will be significant on participants' self-reported germane processing. Specifically, participants in the non-alternating group will report a lower second-stage germane processing load than participants in the alternating group (*H5a*).

H6: Participants' problem-solving performance in the immediate post-test phase will differ across conditions, reflected by problem-solving accuracy and problem-solving time for total test problems. Specifically, participants in the non-alternating group will solve fewer problems and/or solve problems more slowly than participants in the alternating group (*H6a*). The immediate post-test similar question accuracy rate and the immediate post-test transfer question accuracy rate were also evaluated. Participants in the non-alternating group will have a lower accuracy rate than participants in the alternating group in similar questions (*H6b*) and transfer questions (*H6c*).

H7: Participants' problem-solving performance in the delayed post-test phase will differ across conditions, reflected by problem-solving accuracy and problem-solving time for total test problems. Specifically, participants in the non-alternating group will solve fewer problems and/or solve problems more slowly than participants in the alternating group (*H7a*). The delayed post-test similar question accuracy rate and the delayed post-test transfer question accuracy rate were also evaluated. Participants in the non-alternating group will have a lower accuracy rate than participants in the alternating group in similar questions (*H7b*) and transfer questions (*H7c*).

8.1 Method

8.1.1 Sample Size and Participants

Power Analysis. To meet the power requirements of the current research, we kept the parameters of the power analysis as the first experiment. Again, the power analysis was carried out as a one-tailed test, with effect size $d = 0.57$, $\alpha = 0.05$ and desired power = 0.9, which gave a result that at least 54 participants for each group and 108 participants in total were required.

Participants. The third experiment was conducted from September 2024 to October 2024 at East China Normal University (ECNU), Shanghai, China. All participants are ECNU or nearby college students and teachers. A total of 108 participants attended the experiment. One participant dropped out halfway and did not attend the second day's delayed post-test, and the remaining 107 participants completed the whole experiment. All participants were educated to at least the undergraduate level (including those who were currently in undergraduate programs) and were able to read English in paragraphs. All participants were randomly assigned to the control condition (applying a non-alternating finger tracing strategy, N = 54) and the experimental condition (application of an alternating fingers tracing strategy, N = 54). The gender ratio of the participants was not balanced (89 females, 28 males, 1 preferred not to say), but it was consistent with the distribution of the population where the data were collected. Participants' ages ranged from 18 to 35 years old (*M* = 22.26, *SD* = 2.59).

8.1.2 Materials and Procedures

As in Experiment 1 and Experiment 2, recruited participants took part in the experiments individually in a quiet room with guidance from the PhD researcher (Xufei Zhang). Experiment 3 also contained a delayed post-test, which required participants to come back again on the next day to complete it. The paper-based experimental materials were further adjusted based on Experiment 1 and Experiment 2; the full set of experimental materials is given in Appendix D. The whole experiment needed about 35-50 minutes, with the first day lasting 30-40 minutes and the second day lasting 5-10 minutes. The first day's experiment started with the consent phase, followed by the demographics phase, the pre-test phase, the learning and self-reported phase, and the immediate post-test phase. The second day's experiment contained the delayed post-test phase and the debriefing phase. The biggest difference between Experiment 3 and the previous two experiments was that Experiment 3 integrated the learning process with the self-assessment process. The learning and self-report phases were divided into two stages. Participants were requested to learn the first-stage worked examples and then complete a self-report questionnaire based on the first-stage learning. After that, they needed to learn the second-stage work examples and complete the self-report questionnaire again according to the second-stage learning.

Participants in both conditions received identical instructions except for the learning phase, in which they were told to use different tracing strategies (according to their condition). To avoid ambiguity caused by translation, all the materials mentioned above were paper-based and presented in printed English (adapted from materials used in Wang et al., 2022). Though English was not the first language of the participants, English as a compulsory course in the national college entrance exam had been studied by all participants for at least 10 years (from Grade 3 to Grade 12). With years of compulsory English courses, all participants were able to read English in paragraphs. Additional oral instruction between phases was given in the participants' first language (Mandarin). Participants were also encouraged to ask in Mandarin if they had any doubts about the experiment. The whole experiment was videotaped after gaining participants' consent. Recorded participants' tracing actions were counted as a check on treatment fidelity (Mowbray et al., 2003), in line with previous CLT research (Eielts et al., 2020; Galbraith & Ginns, 2023).

Consent Phase. As in Experiment 1 and Experiment 2, participants were notified of the basic information about the experiment and signed a consent form. To avoid the subject-expectancy effect (Rosenthal & Fode, 1963), the current research hid the real purpose of the experiment from participants. Participants were not informed that the experiment was designed to research the tracing strategy. Instead, they were told that this experiment was used to research a new method of two-digit multiplication.

To randomly assign participants to experimental conditions, participants with odd-numbered positions in the experiment sequence were asked to flip a coin to determine their condition assignment. If the coin landed heads up, the participant was assigned to the non-alternating group, and if tails, the participant was assigned to the alternating group.

Participants with even-numbered positions in the experiment sequence did not need to flip a coin, and they were automatically assigned to the opposite condition of the preceding participant.

The video camera was aimed at the desktop portion and only captured participants' hand movements. Video recording commenced after the participant signed the consent form.

Demographics Phase. The experiment officially started with participants answering a survey about their basic demographic information. Their gender and age were collected by choice and fill-in-the-blank questions, along with the same handedness survey (Nicholls et al., 2013) used in Experiments 1 and 2.

Pre-Test Phase. The pre-test phase was designed to measure participants' prior knowledge. The calculation fluency test (Sowinski et al., 2014) used in Experiments 1 and 2 was applied in this phase.

Learning and Self-report Phase. The learning material was similar to that used in Experiment 2, with the following adjustments.

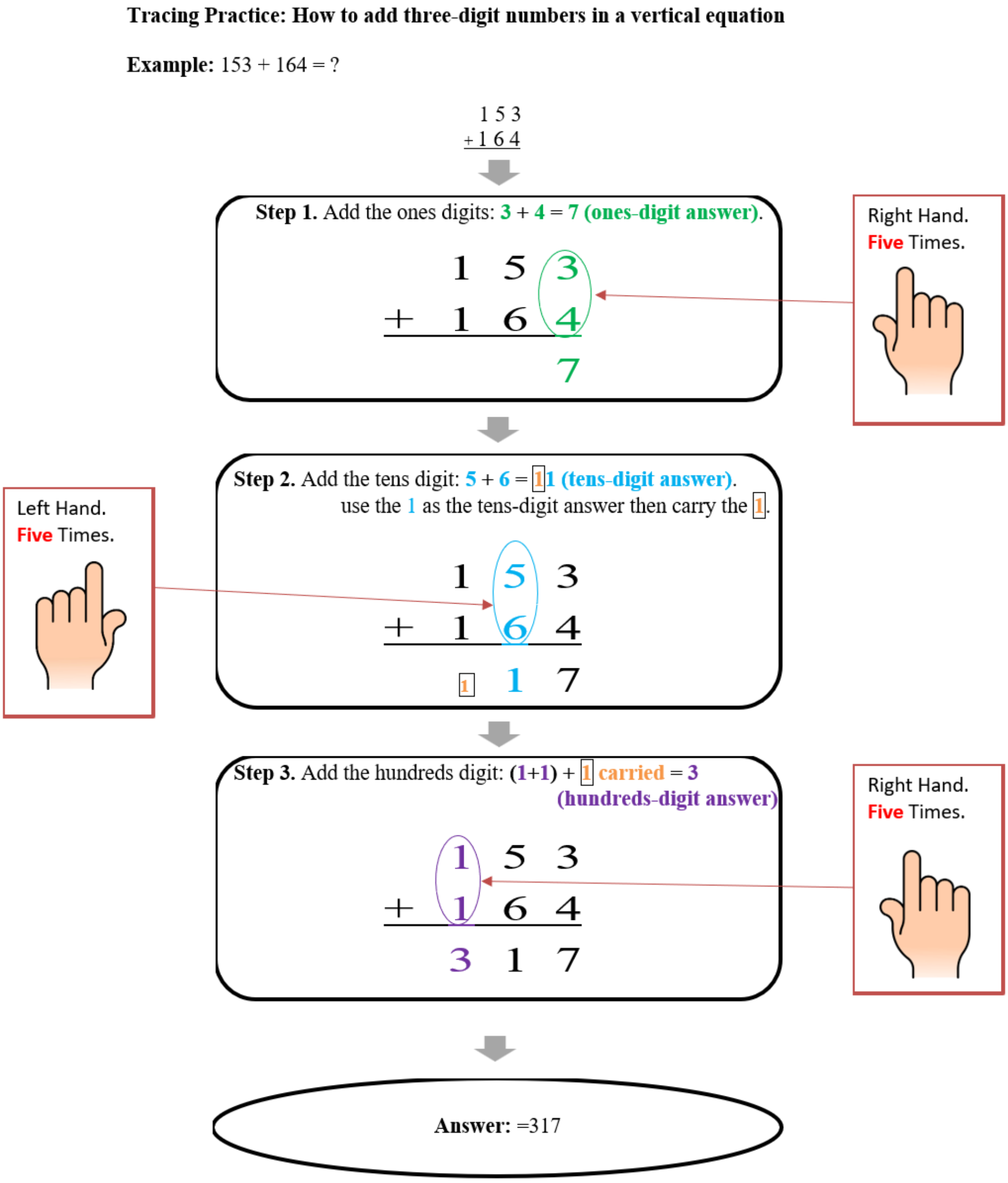
The tracing practice page was redesigned to be more intuitive compared to the tracing practice page used in Experiment 2 (see Figure 10 and Figure 19). In the previous experiments, the tracing practice pages were aimed at giving participants a sense of what finger tracing is and how they should trace the key ellipses with the requested index finger. In Experiment 3, a simple finger picture was used to guide participants' tracing actions, replacing the textual description used in Experiment 2. Because potential split attention effects were detected in previous experiments, the text-based tracing guidance was suspected as a source of interference that prevented participants from distinguishing the key learning points and the learning requirements. The researcher told participants that the meaning of the finger picture was to guide them to use the corresponding hand's index finger to trace the item during the learning process. A more intuitive practice page was intended to help participants understand the tracing guidance more quickly, thereby avoiding their difficulty in distinguishing the learning contents and tracing guidance in the formal learning stage.

When participants practised tracing, they were told that they could ask any questions about the learning requirements based on the practice page, which closely resembles formal learning. They were also informed that, since the formal learning phase would be timed, they would not be able to ask the researcher any questions during this phase. The researcher double-checked with the participants if they totally understood the learning requirements and the words involved in the learning steps before they proceeded to formal learning.

The tracing practice pages in Experiment 3 were identical for the non-alternating condition and the alternating condition, except for the picture of the hand on the side that asked participants to use the index finger of the specific hand to trace the learning material.

Figure 19

Tracing Practice Page Used in Experiment 3



The sequence and numbers of the worked examples used in Experiment 3 were different from Experiment 1 and Experiment 2. Twelve pages of learning instruction containing six worked samples were applied in Experiment 3, replacing eight pages of instruction that contained four worked examples in previous experiments. The twelve pages of instruction were divided into two parts. Each part comprised six pages of instruction, including three worked examples interleaved with three practice questions. All the worked examples were about the same kind of method, but the worked examples in each part were progressively more difficult. The worked examples were adapted from Wang et al. (2022), Experiment 1, and Experiment 2; materials used in the present experiment further adjusted some specific numbers in the worked examples to keep them at the same appropriate difficulty level.

The first stage of learning in Experiment 3 contained three worked examples and three practice questions. Three worked examples were the same as the first, second and third worked examples in Experiment 2. The first worked example did not contain any carry-on calculation during the solving process (see Figure 20), the second worked example contained one carry-on calculation during the solving process (see Figure 21), while the third worked example contained two carry-on calculations in the process (see Figure 22). These three worked examples referred to the difficulty of level one, level two, and level three, as defined in Experiment 2. In contrast to Experiment 2, which included worked examples of four different difficulty levels, Experiment 3 did not contain a level 4 worked example.

As Figure 20, Figure 21, and Figure 22 show, for each worked example, the two-digit multiplication problem was solved in three steps and instructions were provided at each step.

The instruction included a short sentence explaining the step's calculating process, colour-coded horizontal and vertical formulas with key numbers in them, and colour-coded ellipses displaying the key numbers in the vertical formula. The correct answer to the whole worked example is also shown at the end of the three steps.

Another design detail different from previous experiments in Experiment 3 was the presentation of tracing guidance. In Experiment 1, guidance was presented in a separate text box that was located on the left side of the page. In Experiment 2, guidance was presented with the learning steps that were located in the middle of the page. In Experiment 3, the tracing guidance changed from text to images. Instead of putting a long sentence on the side or in the middle of the page to ask participants to trace with the specific hand, Experiment 3 put a specific hand's image on the correct side of the page to tell participants which hand should be used in this step. This adjustment was made because of the potential split attention effect detected in Experiment 1 and Experiment 2. We also found that participants may find it difficult to discern the actual learning content and distinguish it from the tracing guidance. Figure 23, Figure 24, and Figure 25 show the presenting difference between Experiment 1, Experiment 2 and Experiment 3. In each step in Experiment 3, there was a hand with its index finger outlined in red on the side of the page that guided participants to trace the key ellipses with the requested index finger.

Figure 20

A Worked Example Without Carry-on Calculation in Experiment 3

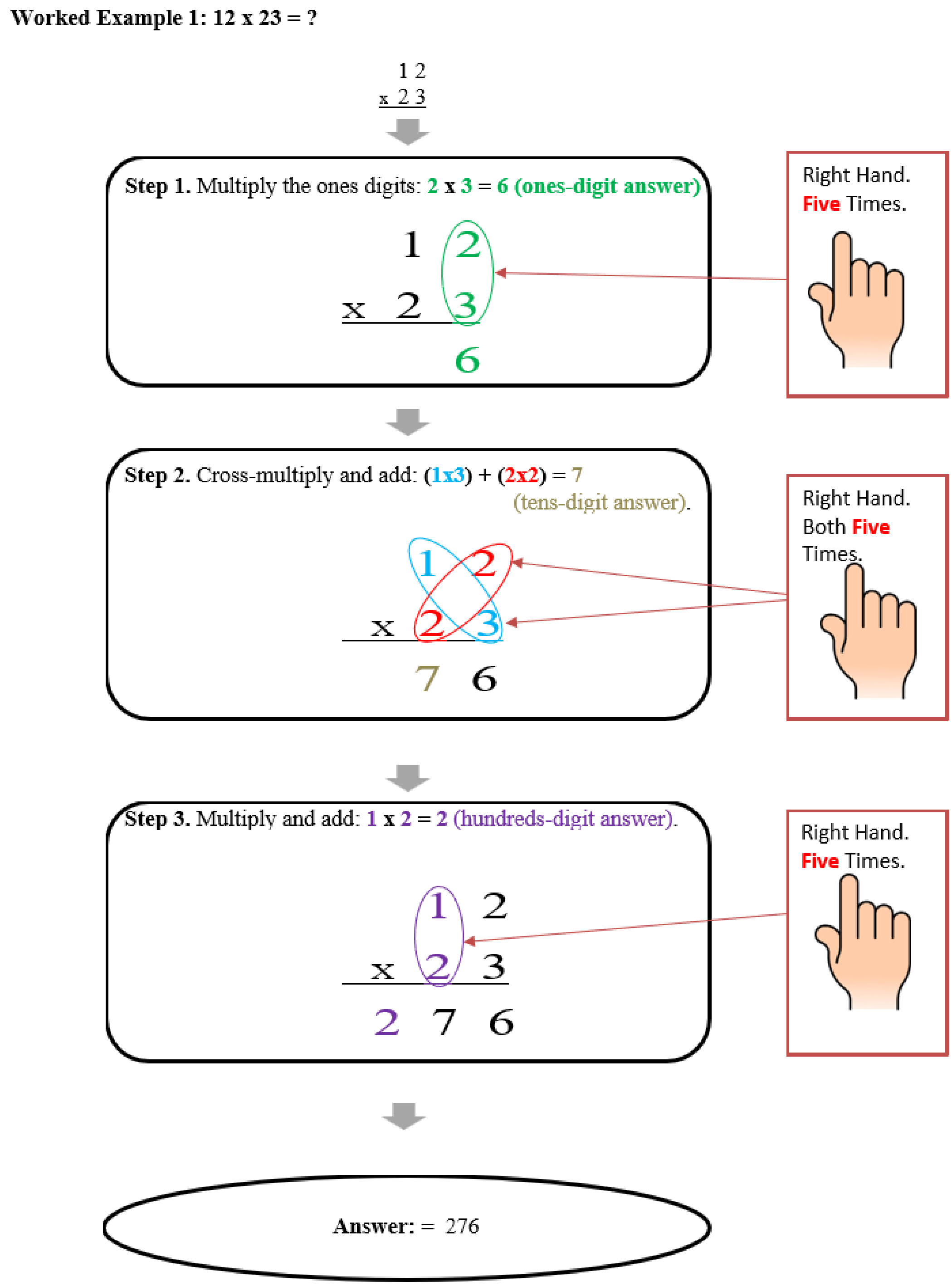


Figure 21

A Worked Example With One Carry-on Calculation in Experiment 3

Worked Example 2: 72 x 54 = ?

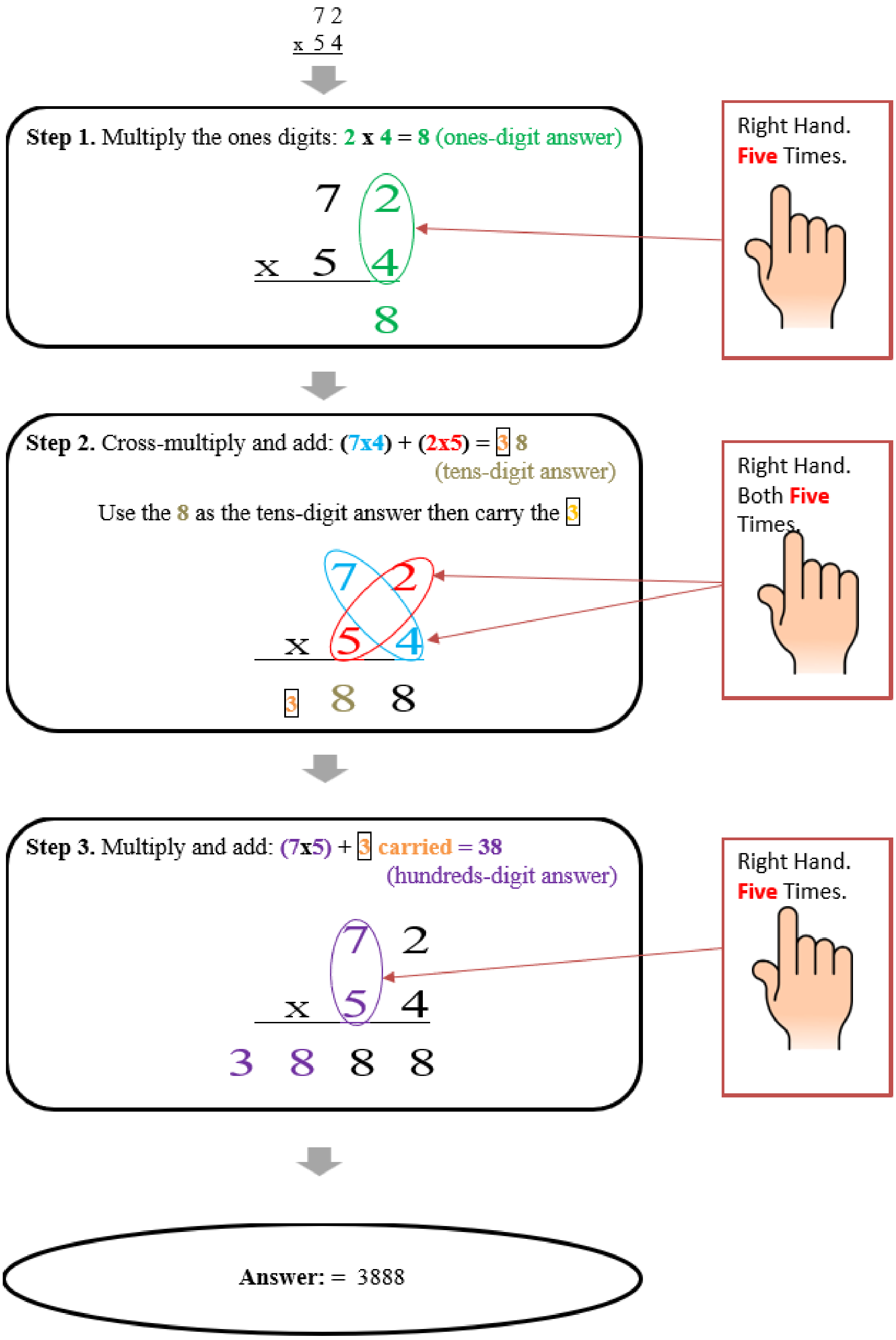


Figure 22

A Worked Example With Two Carry-on Calculations (Simple) in Experiment 3

Worked Example 3: 46 x 23 = ?

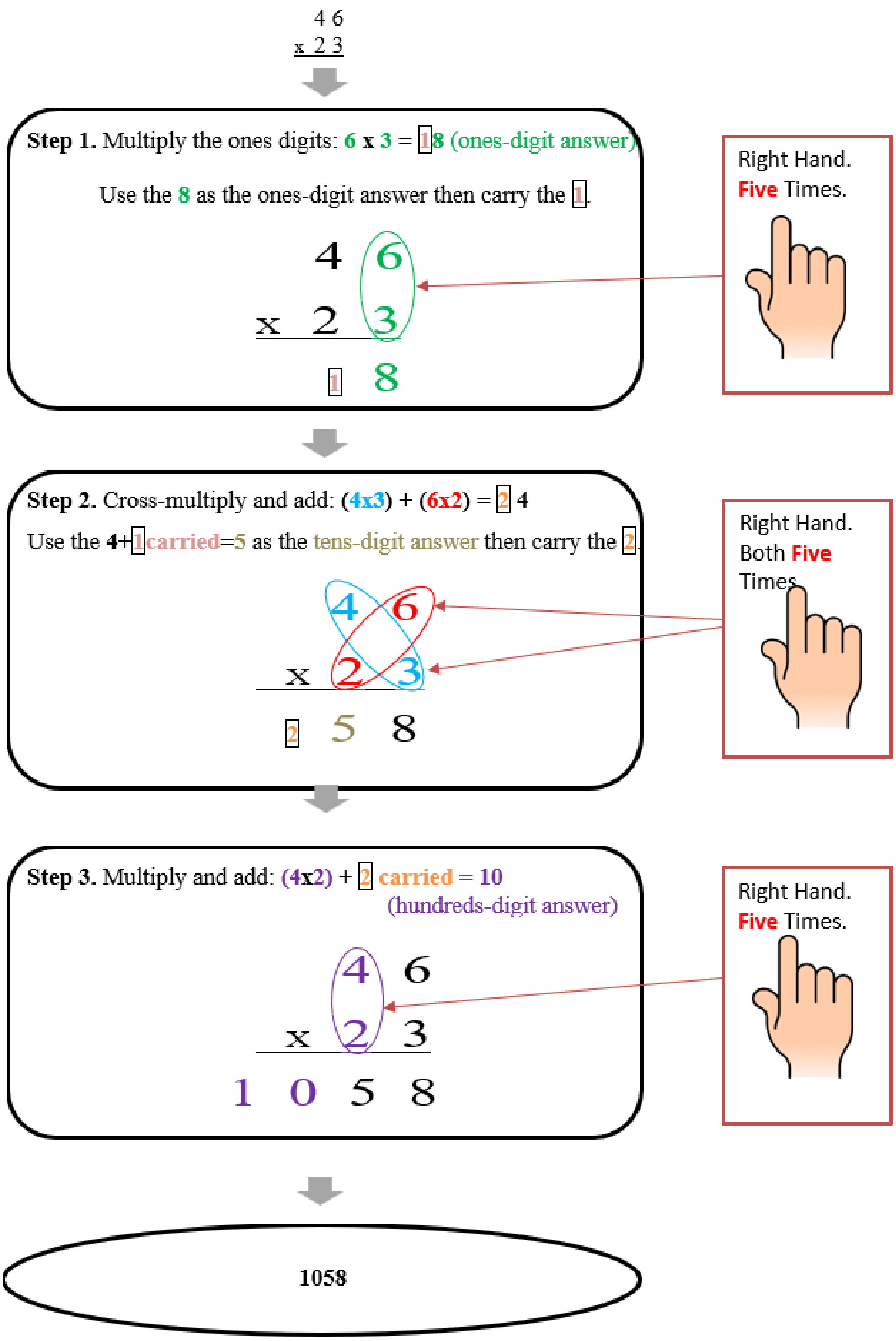


Figure 23

Examples of Steps and Tracing Guidance in Experiment 1

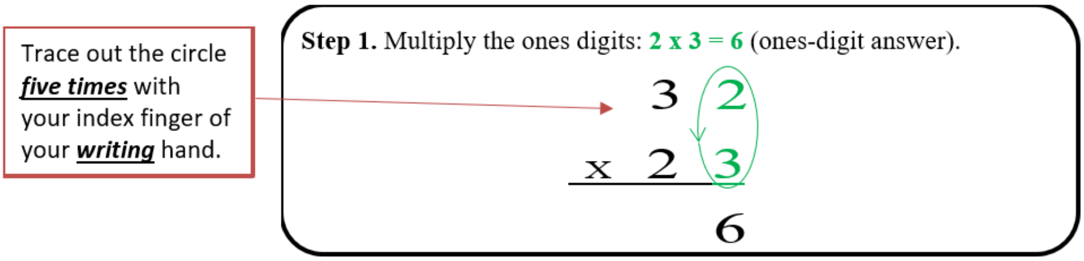


Figure 24

Examples of Steps and Tracing Guidance in Experiment 2

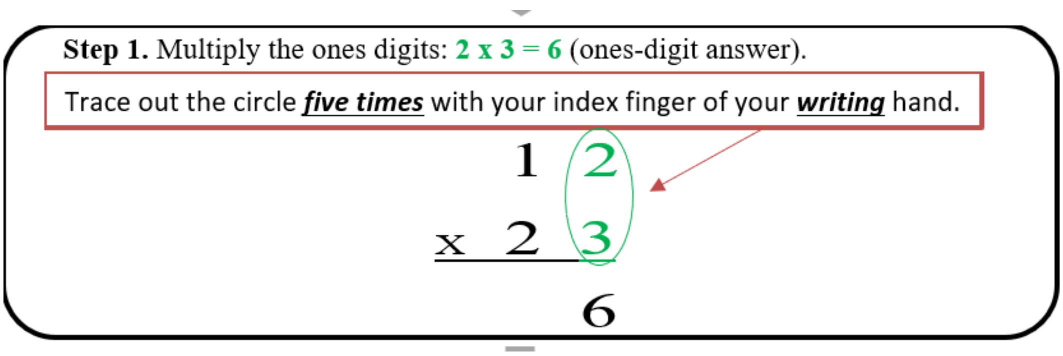
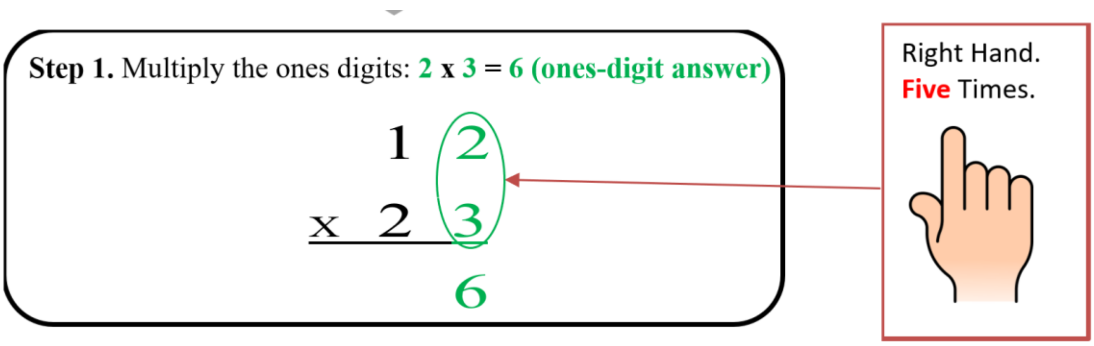


Figure 25

Examples of Steps and Tracing Guidance in Experiment 3



The first stage of learning in Experiment 3 was identical for the non-alternating condition and the alternating condition. Participants in both conditions received the same six-page learning material, in which both of them were requested to use only their right hand's index finger to trace the learning material at all times.

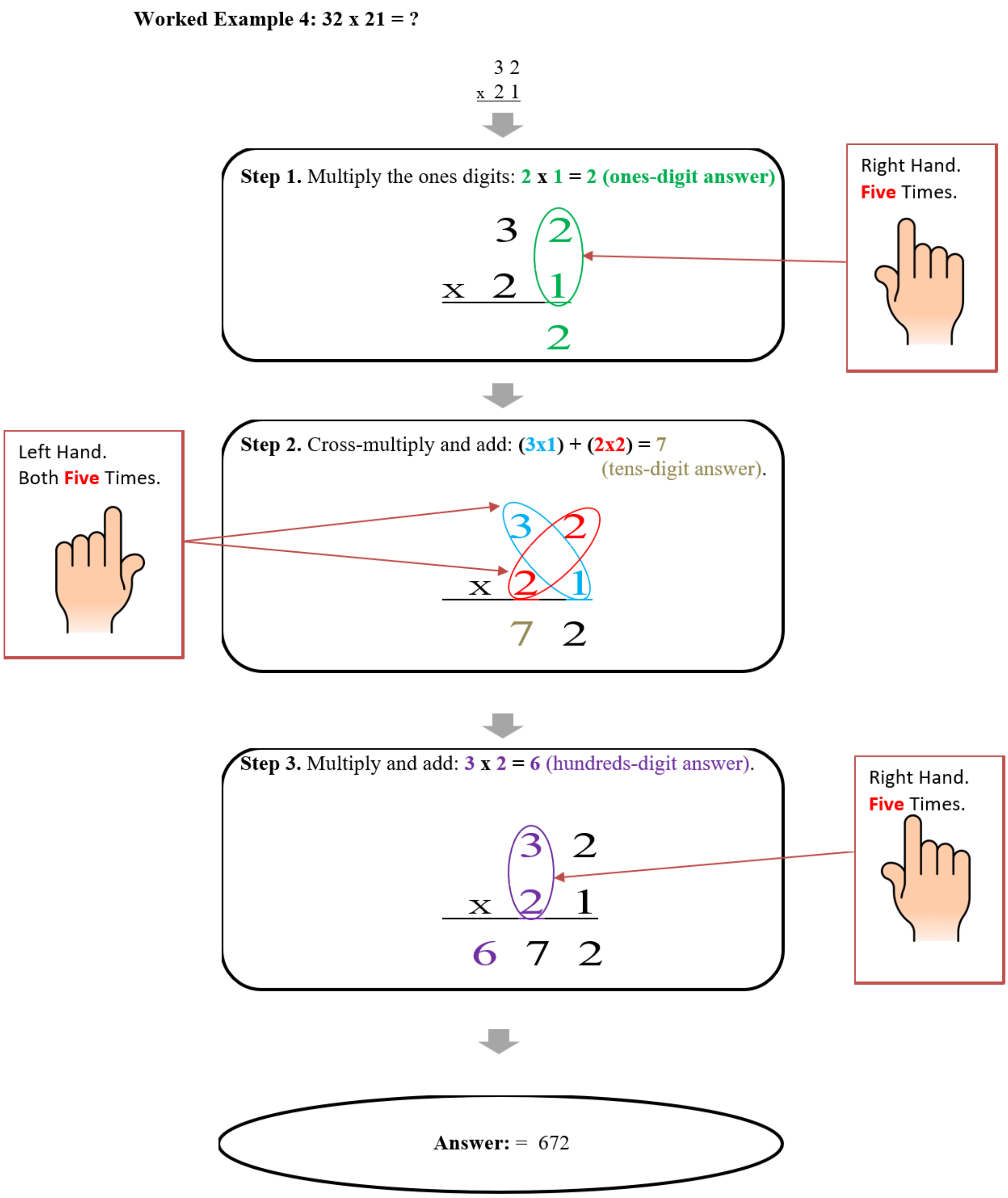
The second stage of learning in Experiment 3 also contained three worked examples (worked example 4, worked example 5, and worked example 6) and three practice questions. These three worked examples had the same difficulty level as the ones in the first stage of learning. Worked example 4 did not contain any carry-on calculation during the solving process (see Figure 20), worked example 5 contained one carry-on calculation during the solving process (see Figure 21), while worked example 6 contained two carry-on calculations in the process (see Figure 22).

The type of tracing guidance applied in the second stage of learning was also in image format, the same as in the first stage of learning. The only difference was that the tracing requirement in the second stage of learning was different for participants in the two conditions. Participants in the non-alternating condition received the learning materials only with the right-hand image beside the learning steps, instructing them to use the right hand's index finger to trace all items. In comparison, participants in the alternating condition received learning materials with right and left-hand images beside the learning steps that requested them to first use the right hand's index finger to trace the first step and then use the left hand's index finger to trace the next step.

Hand images in Experiment 3 were placed on the corresponding side to further help participants quickly understand which hand was to be used for tracing. When they needed to use the index finger of the right hand to trace, the hand image was placed on the right side of the step. When the index finger of the left hand was instructed to be used, the image was presented on the left side. The intention of these changes was to simply and efficiently explain to students what specific action they should take, including which hand to use, thus reducing extraneous load related to the instructions themselves. Figure 26 shows the worked example and its tracing guidance for the alternating condition in the second learning stage in Experiment 3.

Figure 26

The Worked Example for the Alternating Condition in the Second Stage of Learning in Experiment 3



After each stage of learning, participants were asked to complete a twenty-item survey on their motivation and different types of cognitive load. Participants needed to rate each statement from 0 (not at all applicable) to 8 (fully applicable), according to their experience in the learning phase. The intrinsic motivation scale was developed by McAuley et al. (1989), including five different items (e.g., "I enjoyed doing this activity very much"). The cognitive load scale (15 items) was developed by Krieglstein et al. (2023), including five items on intrinsic cognitive load (e.g., "The learning content was difficult to understand"), five items on extraneous cognitive load (e.g., " The design of the learning material made it difficult to find relevant information quickly"), and five items on germane processing (e.g., " I actively reflected upon the learning content"). When some participants asked the precise meaning or the referred-to item of some words and phrases in the statement, the researcher read a pre-prepared standardised translation of the statement in response to such questions from the participants.

One statement regarding germane processing was adjusted in Experiment 3. From previous validity and reliability analysis results in Experiment 1 and Experiment 2, the GCL_2 was found to have low loading for both experiments. The researcher suspected that Chinese participants may have a different understanding of the key term "made an effort to" based on their cultural background. Within the Chinese cultural background, this item may be directly translated to "spent lots of energy to", which emphasised that the task to be completed was not easy and therefore required effort or energy. This term was the key component in GCL_2, which may cause huge misunderstanding if the participants were not able to precisely get its meaning. Thus, the researcher decided to modify the specific expression of GCL_2. The final version of GCL_2 applied in Experiment 3 was formed after the discussion between the current research's author (whose first language was Mandarin and current working language was English) and an Australian local researcher (whose first language and current working language was English). The revised expression of GCL_2 was "I consciously focused to understand the learning content". The other 19 statements did not change, but the translation of the statements was furthermore precise and targeted compared to those applied in Experiment 1 and Experiment 2.

The learning procedure in Experiment 3 also changed compared to Experiment 1 and Experiment 2. In the previous two experiments, participants finished the whole learning phase and then started the self-reported phase. However, in Experiment 3, participants needed to complete the first stage of learning and finish the first stage self-report questionnaire based on the first stage of learning, then they were asked to start the second stage of learning and finish the second stage self-report questionnaire based on the second stage of learning.

At the beginning of this integrated learning and self-reported phase, participants received a two-page instruction telling them how to learn and trace the learning materials in this phase. The researcher also gave oral instructions to help participants understand the procedure and the requirements in this phase. Participants could take their time to read the instructions until they fully understood the requirement, and they were encouraged to ask questions about it. Then, they were asked to trace the ellipses on the tracing practice page with guidance from the researcher. Once they fully understood the learning procedure and requirements, they were allowed to start the formal learning.

Then, participants started to learn the two-digit number multiplication method. They needed to study the three worked examples one by one and answer the practice questions that followed each worked example. Participants had two minutes to study each worked example and 30 seconds to answer the following practice question of the corresponding difficulty. The researcher told participants whether their answer was correct or wrong immediately after participants wrote down the answer. If the answer was wrong, then the participant continued to try until 30 seconds ran out. The sequence was to learn one worked example for two minutes, then answer one question, and learn the next worked example for another two minutes, then answer another question. During the two-minute learning, participants in both conditions traced the ellipses in the learning steps with a specified index finger as the image indicated. The tracing guidance in the two conditions was the same at the first stage of learning. Participants in both groups were asked to use the index finger of the right hand to trace. During participants' learning time, the researcher sat aside and watched participants' learning behaviour, reminding them to focus on learning and use the correct index finger to trace.

After the first stage of learning, participants needed to finish a twenty-item survey immediately to report their motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing based on their learning experiences.

Then they needed to complete the second stage of learning, which also contained three worked examples and three practice questions. They still had two minutes to study each worked example and 30 seconds to answer the following practice question of the corresponding difficulty. The only difference was that the tracing guidance was different to participants in two conditions, which were the writing hand's index finger for the non-alternating tracing condition and alternating index fingers (switching writing and non-writing hands' index fingers) for the alternating tracing condition.

After the second stage of learning, participants completed the twenty-item survey again to report their motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing based on their second stage learning experience.

Immediate Post-Test Phase. Participants answered 20 two-digit multiplication questions in this phase. For each question, they had 20 seconds to answer. They needed to write down the final answer within the time limit. The researcher did not tell the participants if their answers were correct or not. Participants who finished answering one question just moved to the next one, and the researcher started rerecording the time of the next question. The accuracy of participants' answers and the time they spent solving the questions were recorded to reflect their learning performance. Writing an answer that exceeded 20 seconds was recorded as "out of time", and even if the answer was correct, this question was seen as an unsuccessful answer, which was equal to wrong.

Questions in the post-test phase were a little more difficult than the practice questions in the learning phase, with larger numbers and higher difficulty levels. Practice questions were evenly distributed across level 1, level 2 and level 3, with one question for each level as previously described. However, the immediate post-test in Experiment 3 only contained level 3 and level 4 questions. Since Experiment 3 did not include worked examples with difficulty level 4, the level 4 questions in the post-test were seen as transfer questions.

Compared to Experiment 2, which initially only distinguished four levels in the worked examples, the current Experiment 3 intentionally did not include level 4 worked examples in the learning phase but instead incorporated them as part of the transfer questions. Previous research in the cognitive load field has demonstrated that some cognitive load effects may manifest in transfer problems (e.g., Paas & Van Merriënboer, 1994). Thus, Experiment 3 adjusted the design and included transfer questions in the post-tests, with the intention that this could further help the research to explore non-alternating and alternating tracing's learning effects on similar questions and transfer questions. The experiment was originally intended to include 10 similar questions and 10 transfer questions in the immediate post-test, but a design error occurred when choosing the question items. As a result, out of the 20 immediate post-test questions in Experiment 3, 12 were similar questions and 8 were transfer questions.

Delayed Post-Test Phase. The delayed post-test phase was applied one day (approximately 24 hours) after the initial experiment. Participants needed to answer 20 two-digit multiplication questions again in the delayed post-test phase. In contrast to Experiment 1 and Experiment 2, the questions in the delayed post-test were no longer the same as the questions in the immediate post-test. To avoid the potential practice effect (Anastasi, 1988; Campbell & Stanley, 1963), the delayed post-test in Experiment 3 consisted of another 20 questions with similar difficulty as the ones in the immediate post-test. Out of the 20 delayed post-test questions in Experiment 3, 10 were similar questions and 10 were transfer questions.

Debriefing phase. The researcher debriefed the whole experiment to the participants in this phase, clarifying the true purpose of the experiment. Participants were encouraged to express their feelings and thoughts about the experiment and its theoretical assumptions. Some participants asked about their accuracy in the test, and some were interested in the research design. The researcher responded to the participants' questions as much as possible without posing privacy concerns.

8.1.3 Data Processing and Model Check

One hundred and eight participants joined and completed the experiment, and 107 of them completed the experiment in its entirety. These data constituted the full dataset. As participants' behaviours were fully videotaped, the video recordings can be used to check participant compliance with instructions. The fully compliant participants constituted the compliant dataset.

To check participants' compliance, the coding scheme used in Experiment 1 and Experiment 2 was used to quantify participants' effective tracing activities (see Appendix B). Their tracing actions on the required learning materials were recorded as valid counts. Due to the fact that the participants studied six worked examples in Experiment 3, the pass line was adjusted accordingly. Any participant traced less than 96 times in total was deemed "non-compliant" and excluded. Any participant who traced insufficiently (less than 16 times for each worked example) in three or more worked examples was also excluded.

Ninety-seven participants were retained after the compliance check, and 11 participants were excluded. Nine participants were excluded due to insufficient tracing times (participant number 3, 5, 19, 35, 45, 67, 79, 83, 94), and two participants were excluded because of incorrect tracing actions (participant number 1, 44).

The following data analysis was based on the compliant dataset (97 participants' results). The participant who did not come for the delayed post-test was still in this compliant dataset and was included in most analyses, except those analyses based on delayed post-test data.

Handedness Score (Hand Preference and Coordination). As in Experiment 1 and Experiment 2, participants' hand preference was collected via the handedness survey (Nicholls et al., 2013). All participants needed to report their propensity to use one or both hands. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that the handedness survey had acceptable levels of reliability (Cronbach's α = .76; McDonald's ω = .86).

Pre-Test Score (Prior Calculation Ability). As in previous experiments, participants' prior abilities were evaluated via their performance in the calculation fluency test. Participants finished a three-page calculation test on addition, subtraction, and multiplication (one test per page). Every correctly answered question on these pages was counted as one point, producing three separate scores for addition, subtraction, and multiplication fluency.

Table 14 presents participants' performance in the calculation fluency test. On average, participants gained a higher score on addition than subtraction, and both of these scores were further higher than multiplication. The pre-test scores of participants in Experiment 3 were a bit higher than those in Experiment 2, and similar to those in Experiment 1, consistent with Sowinski et al.'s (2014) results. Both Cronbach's alpha (Cronbach, 1951) and McDonald's omega (McDonald, 1999) demonstrated that the pre-test scores had suitable levels of reliability (Cronbach's α = .85; McDonald's ω = .93).

As in Experiment 1 and Experiment 2, though the following learning task in this experiment only involved addition and multiplication in the calculation process, Experiment 3 included subtraction scores together to make the total scores more widely distributed, reflecting participants' calculation ability at a more general level. Scoring the test as the total number of correctly solved items across the three pages was also suggested by the test developer (Sowinski et al., 2014). Thus, three separate scores were added together to generate an overall score to represent their general calculation ability in the following analysis.

In this experiment, participants' prior abilities met the assumption of being normally distributed, based on the Shapiro-Wilk (1965) test. This test was considered the most appropriate method for small sample sizes (<50) and was equally effective compared to other methods for medium-sized samples ($50 \leq n < 300$) (Mishra et al., 2019). As this experiment had 97 participants in total and 48 or 49 participants in each group, the Shapiro-Wilk test was considered a suitable choice for the normality test. The overall distribution (Shapiro-Wilk $W = .99$, $p = .509$) as well as the distribution of each condition satisfied the normal distribution. As Figure 27 shows, the non-alternating tracing group (Shapiro-Wilk $W = .99$, $p = .772$) had a similar distribution to the alternating tracing group (Shapiro-Wilk $W = .98$, $p = .745$).

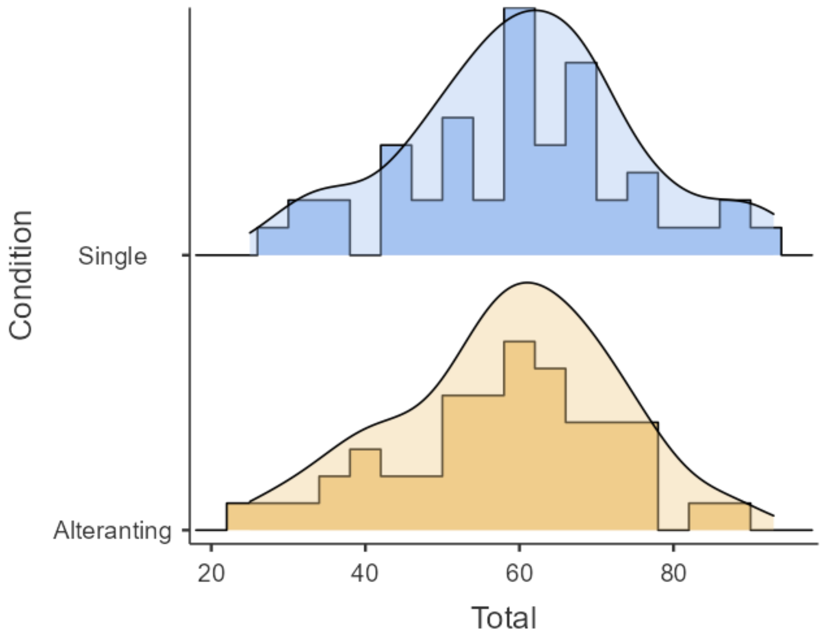
Table 14

Participants' Performance in the Calculation Fluency Test in Experiment 3

Test Type	n	*M*	*SD*
Addition (/60)	97	23.73	6.06
Subtraction (/60)	97	18.82	6.24
Multiplication (/60)	97	16.93	4.20
Total (/180)	97	59.48	14.51

Figure 27

The Distribution of Participants' Total Scores in the Pretest in Experiment 3



Practice Phase Learning Performance. In contrast to Experiment 1 and Experiment 2, Experiment 3 split the worked examples-based learning phase into two stages. The first stage of learning included worked examples 1-3, in which participants in both conditions only used their right index finger to trace. The second stage of learning included worked examples 4-6, in which participants in the non-alternating finger tracing condition only used their right hand's index finger to trace, while participants in the alternating fingers tracing condition alternated their left and right index fingers to trace. Thus, the two stages' learning performance was assessed separately in Experiment 3.

For each stage, participants' learning performance in the learning phase was evaluated via two indicators: the number of practice questions they answered correctly and the time they used to answer the questions. The former was defined as the practice score, and the latter was defined as the practice test time to solution. The practice score was an integer that ranged from 0 to 3. The practice time to solution was obtained by adding up the time participants used to correctly answer each practice question. If the final answer was wrong or participants couldn't give an answer within the time limit (30 seconds), the time of 30 seconds was recorded. Practice time to solution therefore ranged from 0 to 90 seconds.

Post-Lesson Self-Report. Participants' self-reports reflected four variables: intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing, each evaluated by five separate statements. The score ranged from 0 to 8, reflecting participants' agreement with the statement. In contrast to Experiment 1 and Experiment 2, participants completed the self-report survey after each stage of learning in Experiment 3, resulting in each participant having two self-report results regarding their intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing. These results were analysed separately to check their validity and reliability.

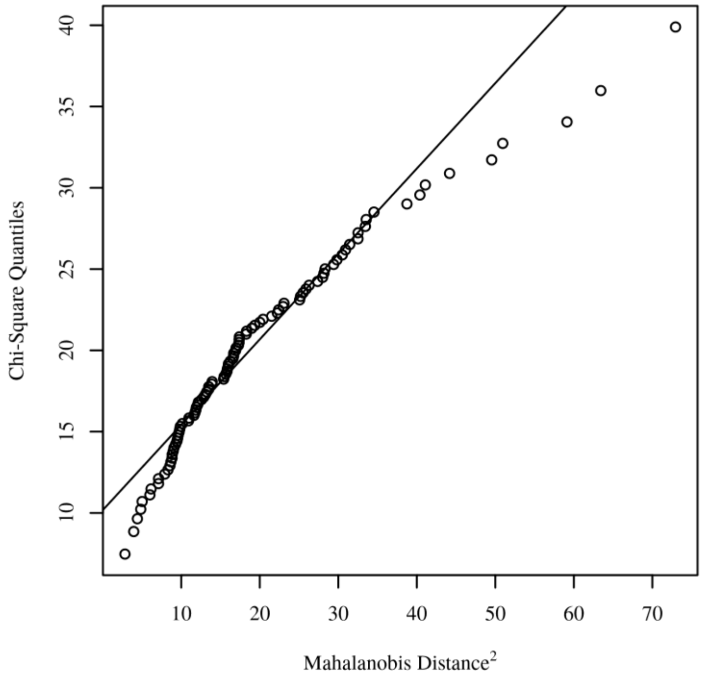
To check the validity of the first-stage survey results, a confirmatory factor analysis was conducted using jamovi (Rosseel et al., 2024). Though CFA's result was generally considered to be reliable only with a large sample size ($n > 300$) (Comrey & Lee, 1992), Wolf et al. (2013) proposed the possibility that a CFA model with high factor loading might be applicable to a minimal sample size lower than 100. Thus, we continued to conduct the CFA model with the current dataset. To check the assumption of multivariate normality, the squared Mahalanobis distances were calculated for the data and plotted against the quantiles of a Chi-square distribution (DeCarlo, 1997). As Figure 28 presents, the assumption of multivariate normality was largely met. Multicollinearity was assessed via the squared multiple correlations and the determinant of the correlation matrix, not finding any variables had an $R^2 > .90$, which indicated low risk in multicollinearity (Kline, 2015).

The Comparative Fit Index (Bentler, 1990), Tucker-Lewis Index (Tucker & Lewis, 1973) and Root Mean Square Error of Approximation (Hu & Bentler, 1999) all showed the model a mediocre fit (CFI = .91, TLI = .90, RMSEA = .09).

Further checking the factor loadings (See Table 15), the indicator "ICL_5A" showed its low loading (< 0.40) again (Cabrera-Nguyen, 2010; Matsunaga, 2010). This indicator was previously found not to be related to the main factor ($p = .122$) in Experiment 1 and had a low loading (= 0.395) in Experiment 2. Now it showed 0.32 in Experiment 3, indicating that this element has been poorly correlated with the main factor. Thus, "ICL_5A" was excluded from the following analysis, and this item should be reconsidered or rephrased in the following research. The model was rerun without this indicator; the results showed a slightly better model fit (CFI = .92, TLI = .90, RMSEA = .09). This indicator was excluded from all the following analyses.

Figure 28

Mahalanobis distance scatterplot testing multivariate normality of the First stage Self-report in Experiment 3



In contrast to Experiment 1 and Experiment 2, the indicator "GCL_2A" was no longer excluded. "GCL_2A" was a low-loading indicator in Experiment 1 and Experiment 2. After rephrasing it in Experiment 3, "GCL_2A" was finally closely related to its main factor, germane processing.

Table 15

Original Factor Loadings of First-stage Self-report in Experiment 3

Factor	Indicator	Estimate	SE	Z	*p*	Stand. Estimate
Intrinsic Motivation	Motivation1A	1.53	0.13	12.02	<.001	0.93
	Motivation2A	1.73	0.12	12.08	<.001	0.93
	Motivation3A	1.88	0.12	12.02	<.001	0.93
	Motivation4A	1.66	0.13	11.81	<.001	0.92
	Motivation5A	1.20	0.20	5.47	<.001	0.52
Intrinsic Cognitive Load	ICL_1A	1.62	0.10	11.86	<.001	0.93
	ICL_2A	1.50	0.09	11.21	<.001	0.90
	ICL_3A	1.79	0.10	11.29	<.001	0.90
	ICL_4AA	1.79	0.12	8.83	<.001	0.77
	ICL_5A	0.92	0.23	3.12	0.00	0.32
Extraneous Cognitive Load	ECL_1A	1.55	0.10	10.00	<.001	0.85
	ECL_2A	1.79	0.10	9.36	<.001	0.81
	ECL_3A	1.71	0.12	7.30	<.001	0.68
	ECL_4A	1.51	0.14	8.21	<.001	0.74
	ECL_5A	1.09	0.17	7.06	<.001	0.66
Germane Processing	GCL_1A	1.49	0.14	7.29	<.001	0.69
	GCL_2A	0.68	0.11	7.31	<.001	0.69
	GCL_3A	1.40	0.13	7.13	<.001	0.67
	GCL_4A	1.55	0.15	6.49	<.001	0.63
	GCL_5A	1.46	0.14	8.60	<.001	0.78

Reliability tests were conducted (Revelle, 2024) to check participants' answers' consistency and dependability. Both Cronbach's Alpha (Cronbach, 1951) and McDonald's Omega (McDonald, 1999) demonstrated that the whole survey and four separate factors had good reliability (See Table 16). Compared to previous experiments, most factors in Experiment 3 had similar reliability in Experiment 2, and had the same or better reliability in Experiment 1.

Participants' motivation and cognitive load were evaluated via the aforementioned 4 factors and 19 items. The average score of each factor was used for the following analyses. All scores ranged from 0 to 8, and higher scores indicate higher motivation or cognitive load.

The Shapiro-Wilk (1965) test was applied to check the normality of participants' average intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing score.

Table 16

Reliability Scores of First-stage Self-report in Experiment 3

Variable	Cronbach's α	McDonald's ω
Intrinsic Motivation (5 items)	.91	.93
Intrinsic Cognitive Load (4 items)	.92	.93
Extraneous Cognitive Load (5 items)	.86	.87
Germane Processing (5 items)	.82	.82

Participants' average first-stage intrinsic motivation (Shapiro-Wilk $W = .96$, $p = .004$) was not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic motivation distribution showed a pronounced left skew: the 25th percentile is 4.8, while the median is as high as 5.8. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic motivation, participant responses of 8, the highest possible response, constituted 20.6%, 21.6%, 18.6%, 16.5%, and 7.2%, for items 1-5 respectively. Based on McHorney and Tarlov's (1995) guidance that ceiling or floor effects are likely to be an issue when more than 15% of respondents respond with either the highest or lowest possible score, a ceiling effect was evident for first-stage intrinsic motivation.

Participants' average first-stage intrinsic cognitive load (Shapiro-Wilk $W = .83$, $p < .001$) was also not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic cognitive load distribution showed a pronounced right skew: the median is 1.0 while the 75th percentile is 1.5. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic cognitive load, participant responses of 0, the lowest possible response, constituted 38.1%, 41.2%, 34.0%, and 40.2% for items 1-4 respectively; thus, a floor effect was evident for self-reports of first-stage intrinsic cognitive load (McHorney & Tarlov, 1995).

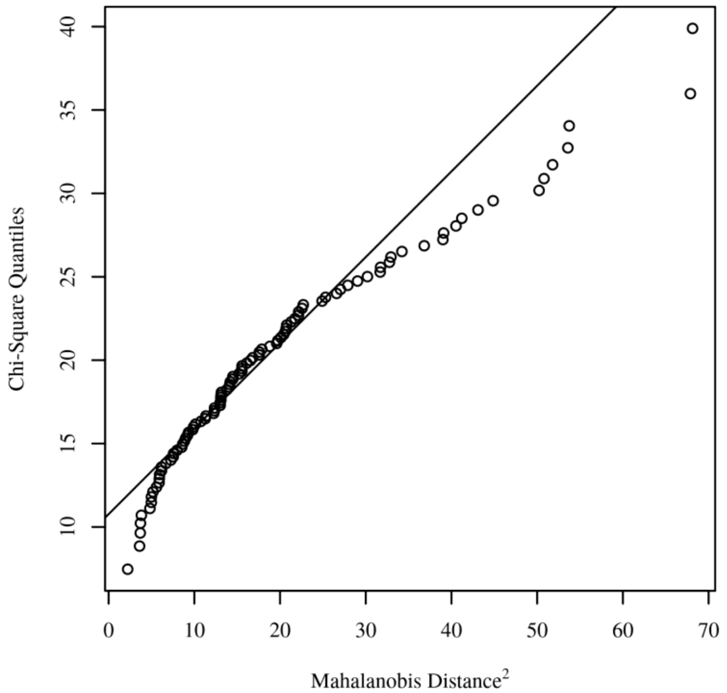
Participants' average first-stage extraneous cognitive load (Shapiro-Wilk $W = .93$, $p < .001$) also were not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the extraneous cognitive load distribution showed a pronounced right skew: the median is 1.2 while the 75th percentile is 2.2. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For extraneous cognitive load, participant responses of 0, the lowest possible response, constituted 23.7%, 32.0%, 29.9%, 37.1% and 36.1%, for items 1-5 respectively; thus, a floor effect was evident for self-reports of first-stage extraneous cognitive load (McHorney & Tarlov, 1995).

Participants' average first-stage germane processing score (Shapiro-Wilk $W = .94$, $p < .001$) also was not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the motivation distribution showed a pronounced left skew: the 25th percentile is 5.4, while the median is as high as 6.4. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For germane processing, participant responses of 8, the highest possible response, constituted 13.4%, 19.6%, 18.6%, 19.6% and 13.4%, for items 1-5 respectively; thus, a ceiling effect was evident for self-reports of first-stage germane processing (McHorney & Tarlov, 1995).

For the second stage self-reported survey, the same validity and reliability tests were conducted. To check the assumption of multivariate normality, the squared Mahalanobis distances were calculated for the data and plotted against the quantiles of a Chi-square distribution (DeCarlo, 1997). As Figure 29 presents, the assumption of multivariate normality was largely met. Multicollinearity was assessed via the squared multiple correlations and the determinant of the correlation matrix; the items "Motivation1B", "Motivation2B" and "Motivation4B" were found to have an $R^2 > .90$, which indicated the high multicollinearity of these items (Kline, 2015).

Figure 29

Mahalanobis Distance Scatterplot Testing Multivariate Normality of Second-stage Self-report in Experiment 3



The results from the confirmatory factor analysis showed a worse model fit than the first stage. Comparative Fit Index (Bentler, 1990), Tucker-Lewis Index (Tucker & Lewis, 1973) and Root Mean Square Error of Approximation (Hu & Bentler, 1999) all showed the model fit with a worse index (CFI = 0.89, TLI = 0.87, RMSEA = 0.11).

Further checking the factor loadings (See Table 17), the factor loadings of indicators were similar to the first stage results. “GCL_2B” was also included in the second stage’s result, and “ICL_5B” still needed to be excluded because of its low loading (Cabrera-Nguyen, 2010; Matsunaga, 2010). The model was rerun without “ICL_5B” and the results showed a similar model fit (CFI = 0.89, TLI = 0.87, RMSEA = 0.11). This indicator was excluded from all the following analyses.

Table 17

Original Factor Loadings of Second-stage Self-report in Experiment 3

Factor	Indicator	Estimate	SE	Z	*p*	Stand. Estimate
Intrinsic Motivation	Motivation1B	1.53	0.13	12.51	<.001	0.95
	Motivation2B	1.73	0.13	12.68	<.001	0.95
	Motivation3B	1.88	0.14	12.42	<.001	0.94
	Motivation4B	1.66	0.15	12.58	<.001	0.95
	Motivation5B	1.20	0.19	7.55	<.001	0.68
Intrinsic Cognitive Load	ICL_1B	1.62	0.09	10.91	<.001	0.89
	ICL_2B	1.50	0.07	11.02	<.001	0.90
	ICL_3B	1.79	0.10	9.05	<.001	0.80
	ICL_4B	1.79	0.12	7.92	<.001	0.74
	ICL_5B	0.92	0.24	3.27	0.00	0.34
Extraneous Cognitive Load	ECL_1B	1.55	0.15	5.22	<.001	0.50
	ECL_2B	1.79	0.12	8.89	<.001	0.77
	ECL_3B	1.71	0.13	11.07	<.001	0.88
	ECL_4B	1.51	0.13	13.76	<.001	1.00
	ECL_5B	1.09	0.21	5.65	<.001	0.53
Germane Processing	GCL_1B	1.49	0.16	7.61	<.001	0.73
	GCL_2B	0.68	0.17	7.55	<.001	0.71
	GCL_3B	1.40	0.13	10.04	<.001	0.87
	GCL_4B	1.55	0.14	8.35	<.001	0.76
	GCL_5B	1.46	0.13	5.25	<.001	0.53

The same reliability tests were conducted (Revelle, 2024) based on the second-stage data. Both Cronbach's Alpha (Cronbach, 1951) and McDonald's Omega (McDonald, 1999) demonstrated that the whole survey and four separate factors had good reliability (See Table 18). Compared to the first-stage results, most factors had higher reliability in the second stage in Experiment 3.

Participants' motivation and cognitive load were evaluated via the aforementioned 4 factors and 19 items. The average score of each factor was used for the following analyses. All scores ranged from 0 to 8, and higher scores indicate higher motivation or cognitive load.

The Shapiro-Wilk (1965) test was applied to check the normality of participants' average intrinsic motivation, intrinsic cognitive load, extraneous cognitive load, and germane processing score.

Table 18

Reliability Scores of Second-stage Self-report in Experiment 3

Variable	Cronbach's α	McDonald's ω
Intrinsic Motivation (5 items)	.95	.96
Intrinsic Cognitive Load (4 items)	.90	.91
Extraneous Cognitive Load (5 items)	.85	.88
Germane Processing (5 items)	.84	.85

Participants' average second-stage intrinsic motivation (Shapiro-Wilk $W = .97$, $p = .014$) was not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic motivation distribution showed a pronounced left skew: the 25^{th} percentile is 4.0 while the median is as high as 5.4. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic motivation, participant responses of 8, the highest possible response, constituted 17.5%, 15.5%, 14.4%, 14.4%, and 7.2%, for items 1-5 respectively. As the second-stage intrinsic motivation had the highest score at 13.8% on average, almost reaching the threshold suggested by McHorney and Tarlov's (1995) guidance that ceiling or floor effects are likely to be an issue when more than 15% of respondents respond with either the highest or lowest possible score.

Participants' average second-stage intrinsic cognitive load (Shapiro-Wilk $W = .79$, $p < .001$) was also not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the intrinsic cognitive load distribution showed a pronounced right skew: the median is 0.5 while the 75^{th} percentile is 1.0. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For intrinsic cognitive load, participant responses of 0, the lowest possible response, constituted 47.4%, 57.7%, 52.6%, and 52.6% for items 1-4 respectively; thus, a floor effect was evident for self-reports of second-stage intrinsic cognitive load (McHorney & Tarlov, 1995).

Participants' average second-stage extraneous cognitive load (Shapiro-Wilk $W = .88$, $p < .001$) also were not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the extraneous cognitive load distribution showed a pronounced right skew: the median is 1.2 while the 75th percentile is 2.2. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For extraneous cognitive load, participant responses of 0, the lowest possible response, constituted 33.0%, 40.2%, 38.1%, 42.3% and 30.9%, for items 1-5 respectively; thus, a floor effect was evident for self-reports of extraneous cognitive load (McHorney & Tarlov, 1995).

Participants' average second-stage germane processing score (Shapiro-Wilk $W = .97$, $p = .024$) also was not normally distributed. As a parameter with a theoretical value ranging from 0 to 8, the quartiles of the motivation distribution showed a pronounced left skew: the 25th percentile is 5.4, while the median is as high as 6.4. To assess the likelihood of ceiling or floor effects, frequency tables for items constituting each of the scales were inspected. For germane processing, participant responses of 8, the highest possible response, constituted 13.4%, 16.5%, 20.6%, 16.5% and 21.6%, for items 1-5 respectively; thus, a ceiling effect was evident for self-reports of germane processing (McHorney & Tarlov, 1995).

Immediate Post-test Phase Learning Performance. Participant's learning performance in the immediate post-test phase was evaluated via two indicators: the number of test questions they answered correctly, and the time they used to answer the questions. The former was defined as the immediate post-test score, and the latter was defined as the immediate post-test time. The immediate post-test score was an integer that ranged from 0 to 20. The immediate post-test time was obtained by adding the time participants used to correctly answer each test question. If a participant could not give a correct answer within the time limit (20 seconds), then the maximum time of 20 seconds was recorded. The immediate post-test time to solution thus ranged from 0 to 400 seconds.

Delayed Post-test Phase Learning Performance. Participants' learning performance in the delayed post-test phase was evaluated via the same indicators as used in the Immediate Post-test. Thus, the delayed post-test score was an integer that ranged from 0 to 20. As for the Immediate Post-test to solution, the Delayed Post-test time to solution was obtained by adding the time participants used to correctly answer each test question. If a participant could not give a correct answer within the time limit (20 seconds), then the maximum time of 20 seconds was recorded. The delayed Post-test time to solution thus ranged from 0 to 400 seconds.

8.1.4 Statistical Analysis

Most procedures and methods to test the hypothesis in Experiment 3 were the same as those in Experiment 1 and Experiment 2. Group equivalence was first checked between the two conditions.

Since the acquisition phase was divided into two stages in Experiment 3, more analyses were applied. The analysis of covariance (ANCOVA) was applied to check the group difference in participants' first-stage practice learning performance, using the prior calculation ability as the covariate. An additional covariate, participants' first-stage practice to solution, was included in the ANCOVA model to represent prior learning ability. Thus, two-covariate ANCOVA was used to check the group difference in participants' second-stage, immediate post-test, and delayed post-test learning performance. Repeated measures ANCOVA (Algina, 1982) was conducted to further analyse the change in participants' mastery level between the immediate post-test and the delayed post-test. Repeated measures analyses were also used to check the group difference in participants' self-reported intrinsic motivation and cognitive load between the two learning stages.

Tests of statistical significance controlled the Type 1 error rate at 0.05.

8.2 Results

8.2.1 Can We Consider Two Groups of Participants as Equivalent?

In this experiment, participants were randomly assigned to conditions, with the expectation that the participants in two conditions would be equivalent in terms of prior ability.

An independent samples *t*-test was conducted to compare participants' prior calculation abilities. As introduced in the earlier section, the pre-test score was normally distributed in both conditions, meeting the normality assumption of the *t*-test. The assumption of homogeneity of variance was also satisfied, $F(1, 95) = 0.06$, $p = .814$. Analysis showed that two conditions did not differ on pre-test scores, $t\ (95) = 0.65$, $p = .516$, $d = 0.13$, 95% CI [-0.27, 0.53], with similar means in the non-alternating condition ($M = 60.46$, $SD = 14.92$) and the alternating condition ($M = 58.53$, $SD = 14.17$). Thus, participants' prior calculation ability was considered equivalent in the two conditions.

Mann-Whitney U test (Mann & Whitney, 1947) was used to analyse participants' handedness across conditions as a supplementary method of the *t*-test, because neither the assumption of normality nor the assumption of homogeneity of variance was met. The result of the Mann-Whitney *U* test was significant, $U = 925.5$, $z = -2.13$, $p = .034$, indicating that participants in the non-alternating condition had lower handedness scores ($M = 0.82$, $SD =$

0.33) than those in the alternating condition (*M* = 0.94, *SD* = 0.13).

Fisher's exact test (Upton, 1992) was used to analyse participants' gender distribution across conditions because of its suitability for small sample sizes, instead of the Pearson chi-square test (Pearson, 1900), which required every cell's frequencies to be bigger than five (Kim, 2016). In the current study, there was an option of "prefer not to say" that was selected by only one participant, which did not meet the basic requirements of the chi-squared test. Fisher's exact test indicated that the non-alternating condition and the alternating condition were not equivalent in terms of the distribution of gender, $p = .020$. More male participants were assigned to the non-alternating condition (12 participants) compared to the alternating condition (four participants), but this difference was not considered to be a substantive concern.

8.2.2 Learning Performance

Did Participants' Prior Abilities Affect Their Learning Performance? Though participants' prior calculation ability knowledge can be considered equivalent at the group level, the results of previous experiments suggested the possible influence of participants' prior calculation ability on learning at the individual level. To control for this potential influence, we need to confirm again whether individuals' calculation abilities were correlated with their learning performance. A correlational analysis was used to investigate the potential relationship between them. As Table 19 shows, participants' pre-test scores were significantly correlated with all types of learning performance. Thus, the pre-test scores were considered as a covariate to avoid the influence of preexisting differences between individuals in the following analyses.

Besides the pre-test score, the practice score (first stage) was also significantly correlated with the immediate post-test score and the delayed post-test score. In Experiment 3, all participants received the same learning material in the first stage of learning. The practice score (first stage) was significantly correlated with all following learning variables, indicating its potential as an additional covariate.

The pre-test score represented participants' calculation ability, so Experiment 1 and Experiment 2 treated the pre-test score as participants' prior abilities. However, the researcher observed some individual cases in previous experiments where some participants performed very well in the pre-test calculations but encountered many difficulties during the worked example learning. Participants' calculation abilities cannot fully represent their ability to learn the specific mental math problem in the current research. Another factor that could represent participants' learning abilities in the current research is therefore desirable.

Table 19

Correlations Between Pre-test Score and Learning Performance in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Pre-test Score (/180)	97	59.49	14.51	—												
2. Practice Score (First Stage) (/3)	97	2.37	0.83	.20*	—											
3. Practice Time to Solution (First Stage) (/90)	97	56.41	16.94	-.38***	-.73***	—										
4. Practice Score (Second Stage) (/3)	97	2.77	0.42	.39***	.18	-.32**	—									
5. Practice Time to Solution (Second Stage) (/90)	97	44.47	11.95	-.45***	-.37***	.57***	-.52***	—								
6. Immediate Post-Test Score (/20)	97	11.43	4.86	.56***	.25*	-.37***	.35***	-.45***	—							
7. Immediate Post-Test Time to Solution (/400)	97	331.55	41.56	-.67***	-.22*	.47***	-.35***	.50***	-.83***	—						
8. Immediate Post-Test Similar Score (/12)	97	7.07	2.71	.61***	.29**	-.40***	.36***	-.49***	.91***	-.83***	—					
9. Immediate Post-Test Transfer Score (/8)	97	4.51	2.35	.54***	.23*	-.42***	.34***	-.45***	.88***	-.82***	.74***	—				
10. Delayed Post-Test Score (/20)	96	13.19	4.61	.54***	.22*	-.29**	.30**	-.38***	.67***	-.63***	.66***	.66***	—			
11. Delayed Post-Test Time to Solution (/400)	96	307.55	50.66	-.73***	-.16	.34***	-.32**	.41***	-.68***	.81***	-.67***	-.69***	-.81***	—		
12. Delayed Post-Test Similar Score (/10)	96	7.34	2.25	.45***	.19	-.23*	.25*	-.31**	.55***	-.48***	.56***	.50***	.89***	-.66***	—	
13. Delayed Post-Test Transfer Score (/10)	96	5.84	2.82	.58***	.21*	-.28**	.29**	-.38***	.67***	-.65***	.63***	.68***	.93***	-.80***	.65***	—

Note. * $p < .05$, ** $p < .01$ *** $p < .001$

Both the practice score (first stage) and the practice time to solution (first stage) can represent distinct forms of relevant prior knowledge in relation to student learning in the second phase. Checking these two factors' correlation with other learning performance factors, the first-stage practice time to solution showed a more significant and higher correlation than the first-stage practice score (see Table 19). In addition, the practice time to solution (first stage) had a more detailed and discriminative distribution than the practice score (first stage). The practice score (first stage) only had four different values: 0, 1, 2, and 3, and inspection of Table 19 reveals the strong likelihood of a ceiling effect on this variate (mean score of 2.37 out of a maximum possible 3 marks). In contrast, the value of the practice time to solution (first stage) ranged from 0 to 90. Thus, the practice time to solution (first stage) was considered more suitable as an additional covariate that represented participants' learning abilities in the context of the mental mathematics strategy.

To apply two covariates in the data analysis process, we need to make sure that these two covariates meet the basic requirements. These two covariates were selected not only because of the correlation index but also because they were logically suitable as indices of participants' prior abilities. The latter reason was particularly important when deciding to conduct a multi-covariate analysis (Tabachnick & Fidell, 2007). The correlation between the pre-test scores and the practice time to solution (first stage) was significant but not strong (Pearson's $r = -0.38$, $p < 0.001$), indicating a low risk of multicollinearity problem (Cohen et al., 2002). Thus, the pre-test scores and the practice time to solution (first stage) were treated as two covariates in the following learning performance analysis.

How Did Participants Perform in the Practice Phase? Since participants' prior abilities were found to correlate with their learning performance, the ANCOVA test (Fox et al., 2024) was used to compare participants' learning performance in the practice phase across the two conditions. Statistical analyses controlled the Type 1 error rate at 0.05. Two indicators, first-stage practice score and first-stage practice time to solution, were separately analysed with the same data processing procedure.

For the first-stage practice score, homogeneity of variances was found via the Levene (1960) test, $F(1, 95) = 0.71$, $p = .403$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.21$, $p < .001$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, referring to Olejnik and Algina’s (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. The interaction effect between condition and prior abilities was not statistically significant, $F(1, 93) = 1.63$, $p = .205$, $\eta^2p = .02$; thus, ANCOVA’s assumption of homogeneity of regression slopes was met. As expected, the main effect of prior calculation ability was statistically significant, $F(1, 94) = 4.01$, $p = .048$, $\eta^2p = .04$; participants with higher pre-test scores also gained higher practice scores. The main effect of condition was not statistically significant, $F(1, 94) = 0.11$, $p = .739$, $\eta^2p = .00$; participants in the non-alternating condition (estimated marginal $M = 2.34$, $SE = .12$) solved problems correctly at a similar rate to those in the alternating condition (estimated marginal $M = 2.40$, $SE = .12$). While no difference between conditions was hypothesised for Stage 1 practice questions, this result should nonetheless be viewed with some caution due to the likely ceiling effect for this variate.

For first-stage practice time to solution, homogeneity of variances was found via

Levene's (1960) test, $F(1, 95) = 3.05$, $p = .084$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.08$, $p = .187$). The interaction effect between the condition and the prior abilities was not found, $F(1, 93) = 0.00$, $p = .996$, showing that the assumption of homogeneity of regression slopes was met. Similar results on the main effects were found. The main effect of prior ability was statistically significant, $F(1, 94) = 15.36$, $p < .001$, $\eta^2 p = .14$; participants with higher pre-test scores spent less time finishing practice questions. The main effect of condition was not found, $F(1, 94) = 0.15$, $p = .700$, $\eta^2 p = .00$; participants in the non-alternating condition (estimated marginal $M = 55.78$, $SE = 2.29$) spent similar time finishing practice questions to those in the alternating condition (estimated marginal $M = 57.03$, $SE = 2.27$).

In summary, aligning with *H1a*, both indicators showed that participants in the non-alternating condition performed the same as those in the alternating condition on the first-stage practice questions.

To analyse participants' second-stage practice performance, a two-covariate ANCOVA test was applied. As introduced in the earlier section, participants' pre-test scores and their first-stage practice time to solution were included as covariates simultaneously to analyse participants' second-stage performance. Statistical analyses controlled the Type 1 error rate at 0.05. Two indicators, second-stage practice score and second-stage practice time to solution, were separately analysed with the same data processing procedure.

For the second-stage practice score, homogeneity of variances was found via the Levene (1960) test, $F(1, 95) = 0.71$, $p = .401$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D =$

0.17, $p < .001$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, referring to Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. No two-way or three-way interaction effects were found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. As expected, the main effect of prior calculation ability was statistically significant, $F(1, 93) = 9.64$, $p = .003$, $\eta^2p = .09$; participants with higher pre-test scores also gained higher second-stage practice scores. The main effect of prior learning ability was marginally significant, $F(1, 93) = 3.89$, $p = .052$, $\eta^2p = .04$; participants who spent less time completing first-stage practice questions gained higher second-stage practice scores. The main effect of the condition was not statistically significant, $F(1, 93) = 0.41$, $p = .525$, $\eta^2p = .00$; participants in the non-alternating condition (estimated marginal $M = 2.80$, $SE = 0.06$) gained similar second-stage practice scores as those in the alternating condition (estimated marginal $M = 2.75$, $SE = 0.06$). However, this result should be viewed with some caution due to the likely ceiling effect for this variate.

For second-stage practice time to solution, the normality of residuals was found via the Lilliefors (1967) test ($D = 0.07$, $p = .199$). However, homogeneity of variances was not found via Levene's (1960) test, $F(1, 95) = 5.16$, $p = .025$, violating the assumption of homoscedasticity. Though the homoscedasticity assumption was not satisfied, the ANCOVA analysis was still applied, referring to Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. No two-way or three-way interaction effects were found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 9.79$, $p = .002$, $\eta^2p = .10$; participants with higher pre-test scores spent less time finishing second-stage practice questions. The main effect of prior learning ability was also statistically significant, $F(1, 93) = 28.90$, $p < .001$, $\eta^2p = .24$; participants who spent less time completing first-stage practice questions spent less time finishing second-stage practice questions. The main effect of condition was not found, $F(1, 93) = 0.73$, $p = .397$, $\eta^2p = .01$; participants in the non-alternating condition (estimated marginal $M = 45.30$, $SE = 1.37$) spent similar time finishing second-stage practice questions to those in the alternating condition (estimated marginal $M = 43.66$, $SE = 1.36$).

Contrary to *H1b*, both indicators showed that participants in the non-alternating condition performed similarly to those in the alternating condition on the second-stage practice questions.

How Did Participants Perform in the Immediate Post-Test Phase? A two-covariate ANCOVA test (Fox et al., 2024) was used to compare participants' learning performance in the immediate post-test in two conditions. Analyses controlled the Type 1 error rate at 0.05. Two indicators, immediate post-test score and immediate post-test time to solution, were separately analysed with the same data processing procedure.

For the immediate post-test score, homogeneity of variances was found via the Levene (1960) test, $F(1, 95) = 0.00$, $p = .987$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.08$, $p = .092$). No two-way or three-way interaction effect was found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 30.91$, $p < .001$, $\eta^2 p = .25$; participants with higher pre-test scores also gained higher immediate post-test scores. The main effect of prior learning ability was statistically significant, $F(1, 93) = 4.49$, $p = .037$, $\eta^2 p = .05$; participants who spent less time completing first-stage practice questions gained higher immediate post-test scores. The main effect of the condition was not statistically significant, $F(1, 93) = 1.43$, $p = .235$, $\eta^2 p = .02$; participants in the non-alternating condition (estimated marginal $M = 10.95$, $SE = 0.57$) gained similar immediate post-test scores to those in the alternating condition (estimated marginal $M = 11.91$, $SE = 0.57$).

For immediate post-test time to solution, homogeneity of variances was found via the Levene (1960) test, $F(1, 95) = 0.33$, $p = .570$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.10$, $p = .027$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, referring to Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. No two-way or three-way interaction effect was found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 51.37$, $p < .001$, $\eta^2 p = .36$; participants with higher pre-test scores spent less time finishing immediate post-test questions. The main effect of prior learning ability was also statistically significant, $F(1, 93) = 10.97$, $p = .001$, $\eta^2 p = .11$; participants who spent less time completing first-stage practice questions spent less time finishing immediate post-test questions. The main effect of condition was not found, $F(1, 93) = 1.78$, $p = .185$, $\eta^2 p = .02$; participants in the non-alternating condition (estimated marginal $M = 335.64$, $SE = 4.31$) spent similar time finishing immediate post-test questions to those in the alternating condition (estimated marginal $M = 327.54$, $SE = 4.27$).

Contrary to *H6a*, both indicators showed that participants in the non-alternating condition performed similarly to those in the alternating condition on the total immediate post-test questions.

To further explore tracing's effect, participants' performance on similar questions and transfer questions in the immediate post-test was separately analysed. As previously noted, the worked examples in Experiment 3 only covered difficulty levels 1 to 3, but level 4 test questions were included in the immediate and delayed post-test. Thus, level 4 questions are considered transfer questions in the current research, in contrast to similar questions. Participants' test score on similar questions and transfer questions was separately analysed following the same procedure as before. Two-covariate ANCOVA tests were applied.

For the similar question score, homogeneity of variances was found via the Levene (1960) test, $F(1, 95) = 3.44$, $p = .067$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.06$, $p = .539$). No two-way or three-way interaction effect was found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 40.22$, $p < .001$, $\eta^2 p = .30$; participants with higher pre-test scores solved more similar questions in the immediate post-test. The main effect of prior learning ability was statistically significant, $F(1, 93) = 5.59$, $p = .020$, $\eta^2p = .06$; participants who spent less time completing first-stage practice questions solved more similar questions in the immediate post-test. The main effect of the condition was not statistically significant, $F(1, 93) = 2.05$, $p = .156$, $\eta^2p = .02$; participants in the non-alternating condition (estimated marginal $M = 6.76$, $SE = 0.30$) solved similar questions in the immediate post-test as similar levels to those in the alternating condition (estimated marginal $M = 7.38$, $SE = 0.30$).

Contrary to *H6b*, participants in the non-alternating condition performed similarly to those in the alternating condition on the similar immediate post-test questions.

For the transfer question score, homogeneity of variances was found via the Levene (1960) test, $F(1, 95) = 0.35$, $p = .555$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.06$, $p = .543$). No two-way or three-way interaction effects were found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 25.18$, $p < .001$, $\eta^2p = .21$; participants with higher pre-test scores solved more transfer questions in the immediate post-test. The main effect of prior learning ability was statistically significant, $F(1, 93) = 8.49$, $p = .004$, $\eta^2p = .08$; participants who spent less time completing first-stage practice questions solved more transfer questions in the immediate post-test. The main effect of condition was marginally statistically significant, $F(1, 93) = 3.07$, $p = .083$, $\eta^2p = .03$; participants in the alternating condition (estimated marginal $M = 4.84$, $SE = 0.27$) tended to solve more transfer questions in the immediate post-test compared to those in the non-alternating condition (estimated marginal $M = 4.16$, $SE = 0.28$).

Consistent with *H6c*, there is tentative evidence that participants in the alternating condition performed better than those in the non-alternating condition on the transfer immediate post-test questions.

How Did Participants Perform in the Delayed Post-Test Phase? A two-covariate ANCOVA test (Fox et al., 2024) was used to compare participants' learning performance in the delayed post-test in two conditions. Analyses controlled the Type 1 error rate at 0.05. Two indicators, delayed post-test score and delayed post-test time to solution, were separately analysed with the same data processing procedure.

For the delayed post-test score, homogeneity of variances was found via the Levene (1960) test, $F(1, 94) = 2.94$, $p = .090$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.08$, $p = .121$). No two-way or three-way interaction effects were found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 92) = 36.59$, $p < .001$, $\eta^2 p = .28$; participants with higher pre-test scores also gained higher delayed post-test scores. However, the main effect of prior learning ability was not significant, $F(1, 92) = 0.93$, $p = .337$, $\eta^2 p = .01$. As the prior learning ability was not effective as a covariate in this ANCOVA model, it was then excluded from the model so as not to lower the power of this analysis (Tabachnick & Fidell, 2014).

One covariate ANCOVA analysis was conducted for the delayed post-test score. Homogeneity of variances was found via the Levene (1960) test, $F(1, 94) = 3.28$, $p = .073$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.07$, $p = .792$). No interaction effect was found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 47.97$, $p < .001$, $\eta^2p = .34$. The main effect of condition was marginally significant, $F(1, 93) = 3.21$, $p = .076$, $\eta^2p = .03$; participants in the alternating condition (estimated marginal $M = 13.88$, $SE = 0.54$) tended to score higher on the delayed post-test than those in the non-alternating condition (estimated marginal $M = 12.50$, $SE = 0.54$).

For the delayed post-test time to solution, homogeneity of variances was found via the Levene (1960) test, $F(1, 94) = 0.01$, $p = .935$, meeting the assumption of homoscedasticity. The normality of residuals was also found via the Lilliefors (1967) test ($D = 0.06$, $p = .452$). No two-way or three-way interaction effects were found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 92) = 84.19$, $p < .001$, $\eta^2p = .48$; participants with higher pre-test scores spent less time finishing delayed post-test questions. However, the main effect of prior learning ability was not significant, $F(1, 92) = 1.09$, $p = .298$, $\eta^2p = .01$. As the prior learning ability was not effective as a covariate in this ANCOVA model, it was then excluded from the model so as not to lower the power of this analysis (Tabachnick & Fidell, 2014).

One covariate ANCOVA analysis was conducted for the delayed post-test time to solution. Homogeneity of variances was found via the Levene (1960) test, $F(1, 94) = 0.00$, $p = .980$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.06$, $p = .848$). No interaction effect was found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 106.63$, $p < .001$, $\eta^2p = .53$.

The main effect of condition was not significant, $F(1, 93) = 0.35$, $p = .558$, $\eta^2 p = .00$; participants in the non-alternating condition (estimated marginal $M = 309.65$, $SE = 5.05$) spent similar time finishing delayed post-test questions to those in the alternating condition (estimated marginal $M = 305.46$, $SE = 5.05$).

Partially consistent with *H7a*, the delayed post-test score showed that participants in the alternating condition tended to perform better than those in the non-alternating condition. However, the delayed post-test time to solution showed that the non-alternating condition performed similarly to those in the alternating condition on the total delayed post-test questions. As the test score was the first priority indicator, *H7a* is tentatively supported in the current experiment.

To further explore tracing's effect, participants' performance on similar questions and transfer questions in the delayed post-test were separately analysed. As with the analysis procedure for the immediate post-test, two-covariate ANCOVA tests were applied.

For similar question scores in the delayed post-test, homogeneity of variances was not found via the Levene (1960) test, $F(1, 94) = 4.57$, $p = .035$, violating the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.09$, $p = .063$). Though the homoscedasticity assumption was not satisfied, the ANCOVA analysis was still applied, referring to Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation. No two-way or three-way interactions were found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 92) = 18.68$, $p < .001$, $\eta^2 p$

= .17; participants with higher pre-test scores also solved more similar questions in the delayed post-test. The main effect of prior learning ability was not significant, $F(1, 92) = 0.60$, $p = .440$, $\eta^2p = .01$. The main effect of the condition was not statistically significant, $F(1, 92) = 2.68$, $p = .105$, $\eta^2p = .03$; participants in the non-alternating condition (estimated marginal $M = 7.01$, $SE = 0.29$) solved a similar number of similar questions in the delayed post-test to those in the alternating condition (estimated marginal $M = 7.68$, $SE = 0.29$).

Contrary to *H7b*, participants in the non-alternating condition performed similarly to those in the alternating condition on the similar delayed post-test questions.

For transfer question scores in the delayed post-test, homogeneity of variances was found via the Levene (1960) test, $F(1, 94) = 0.43$, $p = .512$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.07$, $p = .321$). No two-way or three-way interaction effects were found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 92) = 37.42$, $p < .001$, $\eta^2p = .29$; participants with higher pre-test scores also had a higher transfer question score in the delayed post-test. However, the main effect of prior learning ability was not significant, $F(1, 92) = 0.81$, $p = .370$, $\eta^2p = .01$. As the prior learning ability was not effective as a covariate in this ANCOVA model, it was then excluded from the model so as not to lower the power of this analysis (Tabachnick & Fidell, 2014).

One covariate ANCOVA analysis was conducted for the delayed post-test time to solution. Homogeneity of variances was found via the Levene (1960) test, $F(1, 94) = 0.62$, $p = .434$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.08$, $p = .619$). No interaction effect was found; thus, ANCOVA's assumption of homogeneity of regression slopes was met. The main effect of prior calculation ability was statistically significant, $F(1, 93) = 48.69$, $p < .001$, $\eta^2p = .34$. The main effect of the condition was not significant, $F(1, 93) = 2.34$, $p = .130$, $\eta^2p = .02$; participants in the non-alternating condition (estimated marginal $M = 5.49$, $SE = 0.33$) solved a similar number of transfer questions in the delayed post-test to those in the alternating condition (estimated marginal $M = 6.20$, $SE = 0.33$).

Contrary to *H7c*, participants in the non-alternating condition performed similarly to those in the alternating condition on the transfer delayed post-test questions.

How Did Participants' Learning Performance Change Between the Immediate Post-Test and the Delayed Post-Test? To further analyse the change in participants' mastery level between the immediate post-test and the delayed post-test, a repeated measures ANCOVA (Algina, 1982) was conducted. As applied in the previous ANCOVA analysis, the pre-test score and the first-stage practice time to solution worked as the covariates because they were correlated with participants' immediate post-test performance and delayed post-test performance. Four indicators, the test score, the time to solution, the similar questions accuracy rate, and the transfer questions accuracy rate, were separately analysed with the same data processing procedure. The accuracy rates were calculated by dividing the number of questions answered correctly by the total number of corresponding questions. The accuracy rates were used to analyse participants' learning performance on the similar questions and the transfer questions because the total number of similar and transfer questions was different in the immediate post-test and in the delayed post-test.

For the test score, homogeneity of variances was found via the Levene (1960) test both in the immediate post-test, $F(1, 94) = 1.97$, $p = .163$, and the delayed post-test, $F(1, 94) = 3.68$, $p = .058$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.06$, $p = .048$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

The main effect for test time was significant, $F(1, 92) = 19.25$, $p < .001$, $\eta^2p = .17$, indicating there were significant differences between the immediate post-test score and delayed post-test score; participants gained higher test scores in the delayed post-test (estimated marginal $M = 13.19$, $SE = 0.38$) than in the immediate post-test (estimated marginal $M = 11.46$, $SE = 0.41$). The main effect for pre-test score was significant, $F(1, 92) = 44.66$, $p < .001$, $\eta^2p = .33$, indicating participants with higher prior calculation ability also gained higher test scores. The main effect for first-stage practice time was marginally significant, $F(1, 92) = 3.14$, $p = .080$, $\eta^2p = .03$, indicating participants with higher prior learning ability also gained higher test scores. The main effect for the condition was marginally significant, $F(1, 92) = 2.89$, $p = .093$, $\eta^2p = .03$, indicating participants in the alternating condition gained higher test scores (estimated marginal $M = 12.90$, $SE = 0.48$) than in the non-alternating condition (estimated marginal $M = 11.74$, $SE = 0.48$). The interaction effect between the test time and condition was not significant, $F(1, 92) = 0.38$, $p = .537$, $\eta^2p = .00$, indicating that participants allocated to different conditions had similar trends of change on their test scores during the 24-hour delayed time. The interaction effect between the test time and pre-test scores was also not significant, $F(1, 92) = 0.04$, $p = .853$, $\eta^2p = .00$, indicating that participants with different prior calculation ability had similar trends of change on their test scores during the 24-hour delayed time. The interaction effect between the test time and first-stage practice time to solution was also not significant, $F(1, 92) = 1.43$, $p = .235$, $\eta^2p = .02$, indicating that participants with different prior learning ability have similar trends of change on their test scores during the 24-hour delayed time.

For time to solution, homogeneity of variances was found via the Levene (1960) test both in the immediate post-test, $F(1, 94) = 0.12$, $p = .736$, and the delayed post-test, $F(1, 94) = 0.00$, $p = .991$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.08$, $p = .006$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

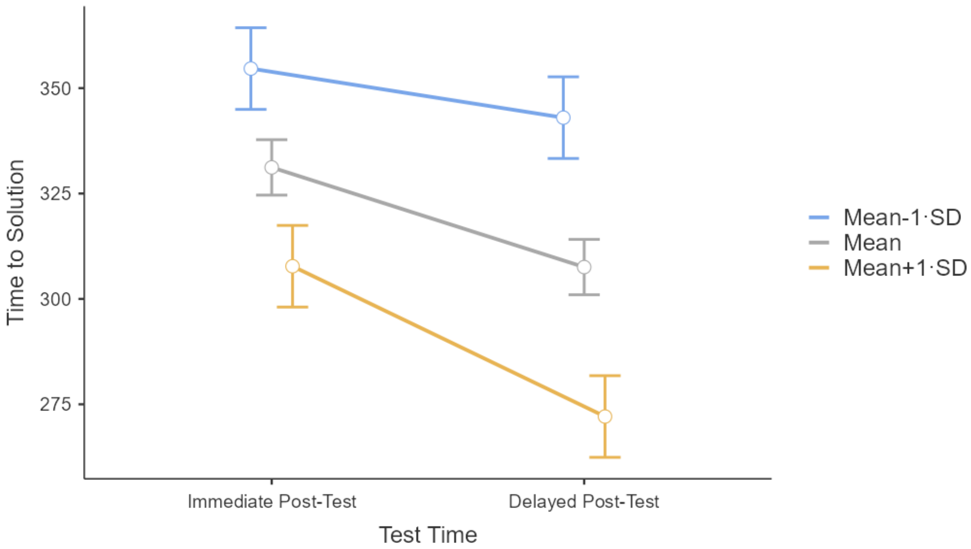
The main effect for the test time was significant, $F(1, 92) = 67.90$, $p < .001$, $\eta^2p = .43$, indicating there were significant differences between the immediate post-test time to solution and delayed post-test time to solution; participants spent less time to completing the test questions in the delayed post-test (estimated marginal $M = 307.55$, $SE = 3.57$) than in the immediate post-test (estimated marginal $M = 331.20$, $SE = 3.06$). The main effect for the condition was not significant, $F(1, 92) = 1.06$, $p = .307$, $\eta^2p = .01$, indicating participants in the non-alternating condition spent similar time to complete post-tests (estimated marginal $M = 322.47$, $SE = 4.25$) to those in the alternating condition (estimated marginal $M = 316.28$, SE

= 4.25). The main effect for pre-test score was significant, $F(1, 92) = 82.00$, $p < .001$, $\eta^2 p = .47$, indicating participants with higher prior calculation ability also gained higher test scores. The main effect for first-stage practice time was significant, $F(1, 92) = 5.22$, $p = .025$, $\eta^2 p = .05$, indicating participants with higher prior learning ability also gained higher test scores. The interaction effect between the test time and condition was not significant, $F(1, 92) = 0.34$, $p = .562$, $\eta^2 p = .00$, indicating that participants in different conditions have similar trends of change in their test time to solution during the 24-hour delayed time.

The interaction effect between the test time and pre-test scores was significant, $F(1, 92) = 15.21$, $p < .001$, $\eta^2 p = .14$, indicating that participants with different prior calculation ability had different trends of change in their test time to solution during the 24-hour delayed time. To further check the specific "turning point" of the interaction effect, a Johnson-Neyman analysis (Johnson & Fay, 1950; Johnson & Neyman, 1936) was applied to pick the exact point (Hayes & Matthes, 2009). Johnson-Neyman analysis treats each value of the moderator as a possible "turning point" where the effect of the focal predictor may shift from non-significance to significance (Krishna, 2016), identifying a specific range of the pre-test score where the effect of test time is significant and non-significant in current research (Hayes & Matthes, 2009). This analysis showed the effect of time was only significant when participants' pre-test total score was higher than 42.68, covering 87.50% (84 participants) of total observations (96 participants). As Figure 30 shows, participants with higher pre-test total scores reduced more in the time completing the test questions in the delayed post-test. Combining the results of the Johnson-Neyman analysis, we can conclude that participants whose pre-test total score was higher than 42.68 spent significantly less time completing the questions in the delayed post-test compared to the immediate post-test.

Figure 30

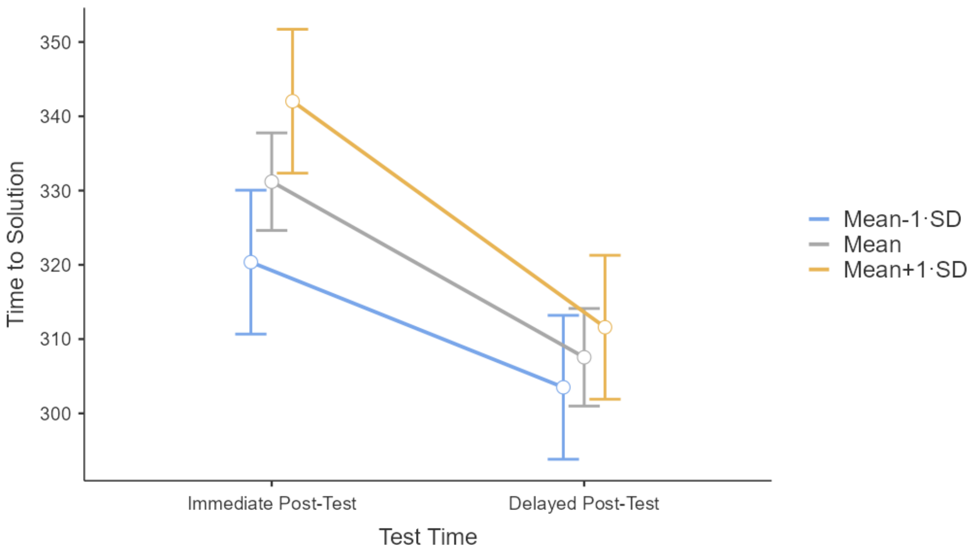
Participants' Time to Solution Under Similar Pre-Test Scores in Immediate Post-Test and Delayed Post-Test in Experiment 3



The interaction effect between the test time and first-stage practice time to solution was also significant, $F(1, 92) = 4.86$, $p = .030$, $\eta^2 p = .05$, indicating that participants with different prior learning ability may have different trends of change in their test time to solution during the 24-hour delayed time. To further check the specific "turning point" of the interaction effect, a Johnson-Neyman analysis (Johnson & Fay, 1950; Johnson & Neyman, 1936) was applied to pick the exact point (Hayes & Matthes, 2009). Johnson-Neyman analysis treats each value of the moderator as a possible "turning point" where the effect of the focal predictor may shift from non-significance to significance (Krishna, 2016), identifying the specific range of the first-stage practice time to solution where the effect of test time is significant and non-significant in current research (Hayes & Matthes, 2009). This analysis showed the effect of time was only significant when participants' first-stage practice time to solution was longer than 27.00, covering 98.96% (95 participants) of total observations (96 participants). As Figure 31 shows, participants with longer first-stage practice time to solution reduced more in the time completing the questions in the delayed post-test. Combining the results of the Johnson-Neyman analysis, we can conclude that participants whose first-stage practice time to solution was longer than 27.00 spent significantly less time completing the questions in the delayed post-test compared to the immediate post-test.

Figure 31

Participant's Time to Solution Under Similar First-Stage Practice Time to Solution in Immediate Post-Test and Delayed Post-Test in Experiment 3



For the similar question's accuracy rate, homogeneity of variances was found via the Levene (1960) test in the delayed post-test, $F(1, 94) = 2.61$, $p = .110$, but not in the immediate post-test, $F(1, 94) = 5.69$, $p = .019$. The normality of residuals was not found via the Lilliefors (1967) test ($D = 0.06$, $p = .060$). Though the homoscedasticity assumption was not satisfied, the ANCOVA analysis was still applied, referring to Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

The main effect for the test time was significant, $F(1, 92) = 45.41$, $p < .001$, $\eta^2p = .33$, indicating there were significant differences on similar question accuracy rate between the immediate post-test and delayed post-test; participants had higher accuracy rate in the delayed post-test (estimated marginal $M = 0.73$, $SE = 0.02$) than in the immediate post-test (estimated marginal *M* = 0.59, *SE* = 0.02). The main effect for the condition was marginally significant, $F(1, 92) = 3.31$, $p = .072$, $\eta^2p = .04$, indicating participants in the alternating condition gained higher test scores (estimated marginal *M* = 0.73, *SE* = 0.02) than in the non-alternating condition post-test (estimated marginal *M* = 0.59, *SE* = 0.02). The main effect for pre-test score was significant, $F(1, 92) = 39.21$, $p < .001$, $\eta^2p = .30$, indicating participants with higher prior calculation ability also gained higher test scores. The main effect for first-stage practice time was marginally significant, $F(1, 92) = 3.23$, $p = .075$, $\eta^2p = .03$, indicating participants with higher prior learning ability also gained higher test scores.

The interaction effect between the test time and condition was not significant, $F(1, 92) = 0.17$, $p = .682$, $\eta^2p = .00$, indicating that participants allocated to different conditions had similar trends of change in their similar question accuracy rate during the 24-hour delayed time. The interaction effect between the test time and pre-test scores was also not significant, $F(1, 92) = 1.23$, $p = .270$, $\eta^2p = .01$, indicating that participants with different prior calculation ability had similar trends of change in their similar question accuracy rate during the 24-hour delayed time. The interaction effect between the test time and first-stage practice time to solution was also not significant, $F(1, 92) = 1.48$, $p = .228$, $\eta^2p = .02$, indicating that participants with different prior learning ability had similar trends of change in their similar question accuracy rate during the 24-hour delayed time.

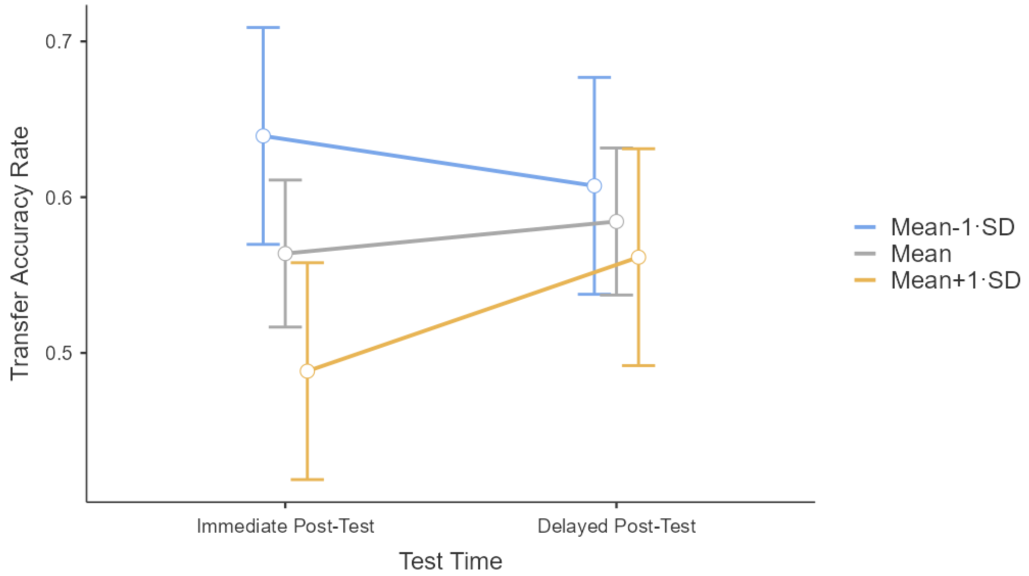
For the transfer question's accuracy rate, homogeneity of variances was found via the Levene (1960) test both in the immediate post-test, $F(1, 94) = 0.08$, $p = .775$, and the delayed post-test, $F(1, 94) = 1.87$, $p = .175$, meeting the assumption of homoscedasticity. The normality of residuals was found via the Lilliefors (1967) test ($D = 0.03$, $p = .930$).

The main effect for the test time was not significant, $F(1, 92) = 0.74$, $p = .391$, $\eta^2p = .01$, indicating there was no significant differences on transfer question accuracy rate between the immediate post-test and delayed post-test; participants had similar accuracy rate in the immediate post-test (estimated marginal $M = 0.56$, $SE = 0.02$) and in the delayed post-test (estimated marginal $M = 0.58$, $SE = 0.02$). The main effect for the condition was marginally significant, $F(1, 92) = 3.45$, $p = .066$, $\eta^2p = .04$, indicating participants in the alternating condition gained higher test scores (estimated marginal $M = 0.61$, $SE = 0.03$) than in the non-alternating condition post-test (estimated marginal $M = 0.54$, $SE = 0.03$). The main effect for pre-test score was significant, $F(1, 92) = 41.16$, $p < .001$, $\eta^2p = .31$, indicating participants with higher prior calculation ability also gained higher test scores. The main effect for first-stage practice time was significant, $F(1, 92) = 4.77$, $p = .032$, $\eta^2p = .05$, indicating participants with higher prior learning ability also gained higher test scores. The interaction effect between test time and condition was not significant, $F(1, 92) = 0.03$, $p = .869$, $\eta^2p = .00$, indicating that participants allocated to different conditions had similar trends of change on their transfer question accuracy rate during the 24-hour delayed time. The interaction effect between the test time and pre-test scores was also not significant, $F(1, 92) = 0.72$, $p = .398$, $\eta^2p = .01$, indicating that participants with different prior calculation ability had similar trends of change on their transfer question accuracy rate during the 24-hour delayed time.

The interaction effect between the test time and first-stage practice time to solution was significant, $F(1, 92) = 4.25$, $p = .042$, $\eta^2p = .04$, indicating that participants with different prior learning ability had different trends of change in their transfer question accuracy rate during the 24-hour delayed time. To further check the specific “turning point” of the interaction effect, a Johnson-Neyman analysis (Johnson & Fay, 1950; Johnson & Neyman, 1936) was applied to pick the exact point (Hayes & Matthes, 2009). Johnson-Neyman analysis treats each value of the moderator as a possible “turning point” where the effect of the focal predictor may shift from non-significance to significance (Krishna, 2016), identifying the specific range of the first-stage practice time to solution where the effect of test time is significant and non-significant in current research (Hayes & Matthes, 2009). This analysis showed that the effect of time was only significant when participants’ first-stage practice time to solution was longer than 68.93, covering 27.08% (26 participants) of total observations (96 participants). As Figure 32 shows, participants with longer first-stage practice time to solution gained more positive progress on the transfer question accuracy rate in the delayed post-test. Combining the results of the Johnson-Neyman analysis, we can conclude that participants whose first-stage practice time to solution was longer than 68.93 had significantly higher transfer question accuracy rate in the delayed post-test compared to the immediate post-test.

Figure 32

Participants' Transfer Question Accuracy Rate Under Similar First-Stage Practice Time to Solution in Immediate Post-Test and Delayed Post-Test in Experiment 3



8.2.3 Post-Lesson Self-Report

Did Participants' Prior Abilities Affect Their Motivation and Cognitive Load?

Similar to Experiment 1 and Experiment 2, a correlational analysis was conducted to determine prior ability's relation with self-reports of intrinsic motivation and cognitive load. Participants' motivation and cognitive load in the first stage and the second stage were separately analysed.

Table 20 presents detailed results of correlational analyses between participants' pre-test scores and first-stage motivation and cognitive load. No intrinsic motivation or any kind of cognitive load was found to be significantly correlated with pre-test scores, indicating we did not need to consider prior calculation abilities in the following analysis on first-stage motivation and cognitive load.

Since two covariates, pre-test score and first-stage practice time to solution, were included in the analysis of participants' second-stage learning performance as prior calculation ability and learning ability, it was necessary to consider these two factors as potential covariates for participants' second-stage motivation and cognitive load. Both the practice score (first stage) and the practice time (first stage) can represent participants' specific learning abilities in the current research. Thus, the correlations between participants' pre-test scores, first-stage practice scores, first-stage practice time to solution, second-stage intrinsic motivation and cognitive load were estimated.

Table 21 presents detailed results of the correlation between participants' pre-test scores, first-stage learning performance and second-stage motivation and cognitive load. The first-stage practice score was found to correlate with participants' second-stage intrinsic cognitive load, suggesting it needs to be considered as a covariate in the following analysis.

Table 20

Correlation Between Prior Calculation Ability, First-stage Motivation, and Cognitive Load in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5
1. Pre-test Score (/180)	97	59.49	14.51	—				
2. Intrinsic Motivation (/8)	97	5.69	1.48	.17	—			
3. Intrinsic Cognitive Load /8)	97	1.06	1.14	-.06	-.29***	—		
4. Extraneous Cognitive Load (/8)	97	1.40	1.13	-.09	-.39***	.76***	—	
5. Germane Processing (/8)	97	6.23	1.09	.14	.64***	-.57***	-.61***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 21

Correlation Between Prior Calculation Ability, Prior Learning Ability, Second-stage Motivation, and Cognitive Load in Experiment 3

Variable	n	M	SD	1	2	3	4	5	6	7
1. Pre-test Score (/180)	97	59.49	14.51	—						
2. Practice Score (First Stage) (/4)	97	2.37	0.83	.20*	—					
3. Practice Time to Solution (First Stage) (/120)	97	56.41	16.94	-.38***	-.73***	—				
4. Intrinsic Motivation (/8)	97	5.24	1.73	.15	-.14	.16	—			
5. Intrinsic Cognitive Load (/8)	97	0.80	1.00	.02	-.24*	.16	-.02	—		
6. Extraneous Cognitive Load (/8)	97	1.46	1.36	-.05	.01	.01	-.17	.44***	—	
7. Germane Processing (/8)	97	6.04	1.22	.10	-.10	.08	.52***	-.17	-.22*	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

As the learning phase was divided into two stages and all participants applied non-alternating tracing in the first stage, the repeated measure analyses were used to analyse participants' intrinsic motivation and cognitive load in Experiment 3. According to the correlation analyses conducted before, the 2 x 2 ANOVA analyses would be applied to check participants' intrinsic motivation, extraneous cognitive load, and germane processing. The 2 x 2 ANCOVA analyses would be applied to check participants' intrinsic cognitive load, setting the first-stage practice score as the covariate.

How Did Participants' Intrinsic Motivation Change Between the First Stage of Learning and the Second Stage of Learning? To analyse the change of participants' intrinsic motivation between the first stage of learning and the second stage of learning, a repeated measures ANOVA (Algina, 1982) was conducted. Learning stage was the within-subject factor in the analysis. The key factor of this analysis was the condition and the interaction effect between the condition and the learning stage.

Homogeneity of variances was found via the Levene (1960) test both in the first stage, $F(1, 95) = 0.10$, $p = .751$, and the second stage, $F(1, 95) = 0.34$, $p = .559$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.07$, $p = .041$). Though the normality assumption was not satisfied, the ANOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

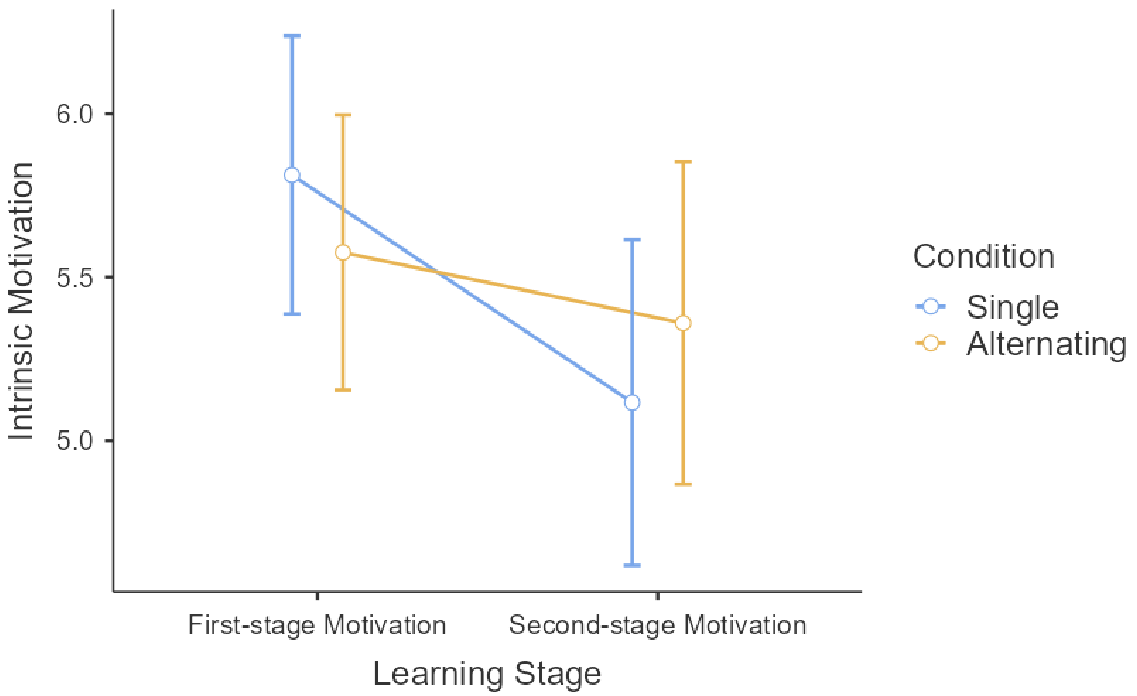
The main effect for the condition was not significant, $F(1, 95) = 0.00$, $p = .993$, $\eta^2 p = .00$, indicating there was no significant general difference between the non-alternating condition and the alternating condition.

The main effect for the learning stage was significant, $F(1, 95) = 15.75$, $p < .001$, $\eta^2 p = .14$, indicating there were significant differences between the first-stage intrinsic motivation and the second-stage intrinsic motivation; participants reported higher intrinsic motivation in the first stage (estimated marginal *M* = 5.69, *SE* = 0.15) than in the second stage (estimated marginal *M* = 5.24, *SE* = 0.18).

The interaction effect between the learning stage and condition was also significant, $F(1, 95) = 4.35$, $p = .040$, $\eta^2 p = .04$, indicating that participants allocated to different conditions have different trends of change on intrinsic motivation between the two learning stages. As Figure 33 shows, participants in the non-alternating condition had a more substantial decline in their intrinsic motivation across the two learning stages compared to those in the alternating condition.

Figure 33

Participants' Intrinsic Motivation in Two Conditions in the First-Stage Learning and the Second-Stage Learning in Experiment 3



How Did Participants' Intrinsic Cognitive Load Change Between the First Stage of Learning and the Second Stage of Learning? To further analyse the change of participants' intrinsic cognitive load between the first stage of learning and the second stage of learning, a repeated measures ANCOVA (Algina, 1982) was conducted. The first-stage practice score worked as the covariate because it was correlated with participants' second-stage intrinsic cognitive load. Learning stage was the within-subject factor in the analysis. The key factor of this analysis was the condition and the interaction effect between the condition and the learning stage.

Homogeneity of variances was found via the Levene (1960) test both in the first stage, $F(1, 95) = 0.50$, $p = .481$, and the second stage, $F(1, 95) = 0.02$, $p = .883$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.11$, $p < .001$). Though the normality assumption was not satisfied, the ANCOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANCOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

The main effect for the condition was not significant, $F(1, 94) = 0.97$, $p = .328$, $\eta^2p = .01$, indicating there was no significant general difference in participants' intrinsic cognitive load between the non-alternating condition and the alternating condition.

The main effect for the learning stage was significant, $F(1, 94) = 14.00$, $p < .001$, $\eta^2p = .13$, indicating there were significant differences between the first-stage intrinsic cognitive load and the second-stage intrinsic cognitive load; participants reported higher intrinsic cognitive load in the first stage (estimated marginal *M* = 1.06, *SE* = 0.11) than in the second stage (estimated marginal *M* = 0.80, *SE* = 0.10). The interaction effect between the learning stage and condition was not significant, $F(1, 94) = 0.28$, $p = .601$, $\eta^2 p = .00$, indicating that participants allocated to different conditions had similar trends of change on intrinsic cognitive load between the two learning stages.

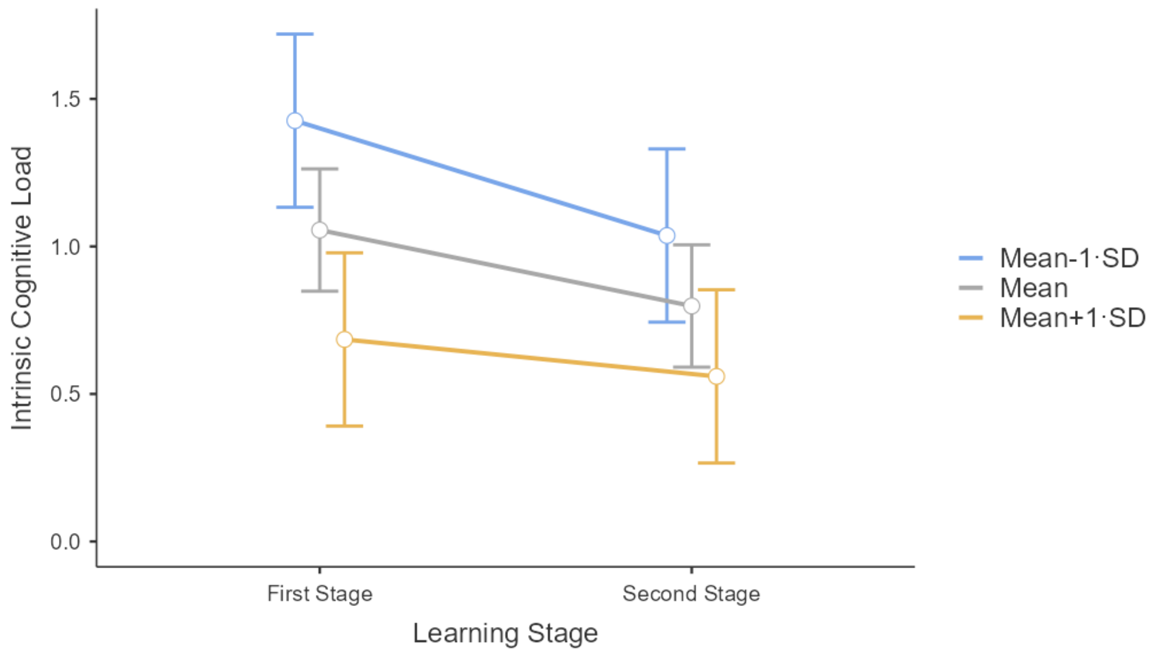
The interaction effect between the learning stage and the condition was not significant, $F(1, 94) = 0.28$, $p = .601$, $\eta^2 p = .00$, indicating that participants in different conditions had similar trends of change on intrinsic cognitive load between the two learning stages.

The interaction effect between the learning stage and the first-stage practice score was marginally significant, $F(1, 94) = 3.66$, $p = .059$, $\eta^2 p = .04$, indicating that participants with different first-stage practice scores may have had different trends of change on intrinsic cognitive load between the two learning stages. Though a marginal significance was detected in the analysis, the Johnson-Neyman analysis (Johnson & Fay, 1950; Johnson & Neyman, 1936) was still applied to further check if there was a specific "turning point" of the interaction effect (Hayes & Matthes, 2009). Johnson-Neyman analysis treats each value of the moderator as a possible "turning point" where the effect of the focal predictor may shift from non-significance to significance (Krishna, 2016), identifying a specific range of the first-stage practice score where the effect of the learning stage is significant and non-significant in current research (Hayes & Matthes, 2009). This analysis showed that the effect of the learning stage was only significant when participants' first-stage practice score was lower than 2.95, covering 43.30% (42 participants) of total observations (97 participants). As

Figure 34 shows, participants with lower first-stage practice scores reduced more on the intrinsic cognitive load in the second stage. Combining the results of the Johnson-Neyman analysis, we can conclude that participants whose first-stage practice score was lower than 2.95 showed a significant decline in their intrinsic cognitive load during the two learning stages.

Figure 34

Participants' Intrinsic Cognitive Load Under Similar First-Stage Practice Score in the First-Stage Learning and the Second-Stage Learning in Experiment 3



How Did Participants' Extraneous Cognitive Load Change Between the First Stage of Learning and the Second Stage of Learning? To further analyse the change in participants' extraneous cognitive load between the first stage of learning and the second stage of learning, a repeated measures ANOVA (Algina, 1982) was conducted. Learning stage was the within-subject factor in the analysis. The key factor of this analysis was the condition and the interaction effect between the condition and the learning stage.

Homogeneity of variances was found via the Levene (1960) test both in the first stage, $F(1, 95) = 0.07$, $p = .799$, and the second stage, $F(1, 95) = 1.22$, $p = .272$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.10$, $p < .001$). Though the normality assumption was not satisfied, the ANOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

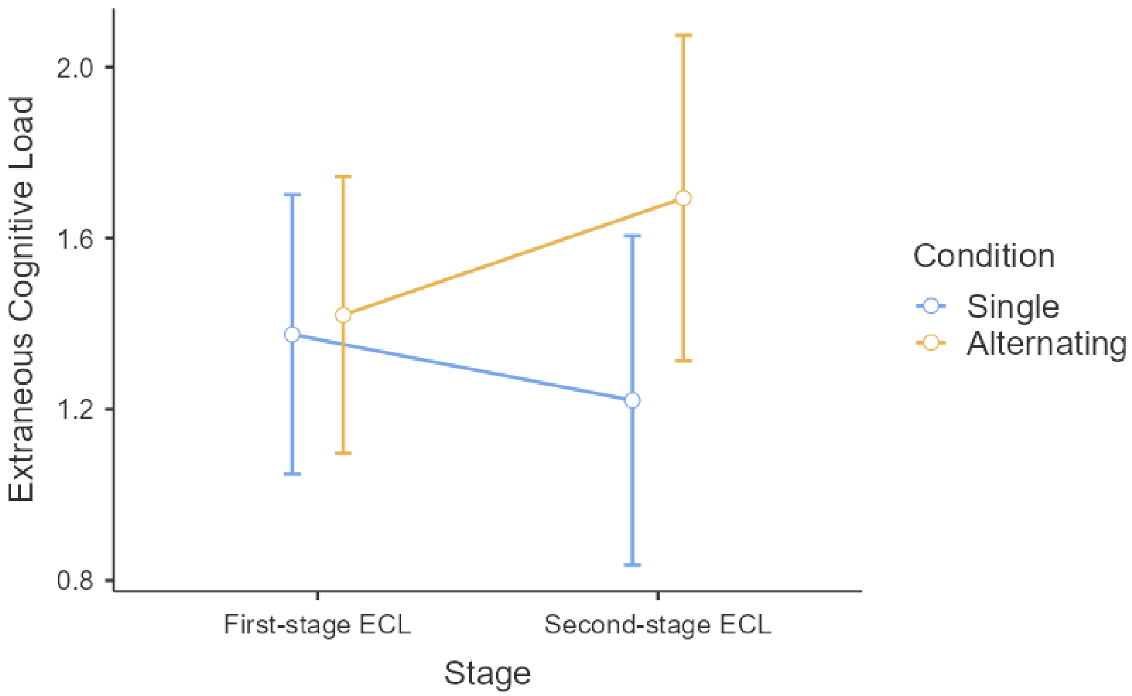
The main effect for the condition was not significant, $F(1, 95) = 1.35$, $p = .248$, $\eta^2p = .01$, indicating there was no significant general difference in participants' extraneous cognitive load between the non-alternating condition and the alternating condition.

The main effect for the learning stage was not significant, $F(1, 95) = 0.25$, $p = .619$, $\eta^2p = .00$, indicating there was no significant difference between the first-stage extraneous cognitive load and the second-stage extraneous cognitive load; participants reported similar extraneous cognitive load in the first stage (estimated marginal $M = 1.40$, $SE = 0.12$) to those in the second stage (estimated marginal $M = 1.46$, $SE = 0.14$).

The interaction effect between the learning stage and condition was marginally significant, $F(1, 95) = 3.20$, $p = .077$, $\eta^2 p = .03$, indicating that participants allocated to different conditions may have different trends of change in extraneous cognitive load between the two learning stages. As Figure 35 shows, participants in the non-alternating condition had lower extraneous cognitive load in the second stage of learning compared to the first stage of learning, while participants in the alternating condition had higher extraneous cognitive load in the second stage of learning compared to the first stage of learning.

Figure 35

Participants' Extraneous Cognitive Load in Two Conditions in the First-Stage Learning and the Second-Stage Learning in Experiment 3



How Did Participants' Germane Processing Change Between the First Stage of Learning and the Second Stage of Learning? To further analyse the change of participants' extraneous cognitive load between the first stage of learning and the second stage of learning, a repeated measures ANOVA (Algina, 1982) was conducted. Learning stage was the within-subject factor in the analysis. The key factor of this analysis was the condition and the interaction effect between the condition and the learning stage.

Homogeneity of variances was found via the Levene (1960) test both in the first stage, $F(1, 95) = 0.85$, $p = .359$, and the second stage, $F(1, 95) = 1.10$, $p = .297$, meeting the assumption of homoscedasticity. However, the normality of residuals was not found via the Lilliefors (1967) test ($D = 0.09$, $p < .001$). Though the normality assumption was not satisfied, the ANOVA analysis was still applied, based on Olejnik and Algina's (1984) findings that ANOVA demonstrated robustness when either the assumption of normality or the assumption of homoscedasticity was violated in isolation.

The main effect for the condition was not significant, $F(1, 95) = 0.04$, $p = .850$, $\eta^2p = .00$, indicating there was no significant general difference in participants' germane processing between the non-alternating condition and the alternating condition.

The main effect for the learning stage was not significant, $F(1, 95) = 2.41$, $p = .124$, $\eta^2p = .03$, indicating there was no significant difference between the first-stage germane processing and the second-stage germane processing; participants reported similar germane processing in the first stage (estimated marginal $M = 6.23$, $SE = 0.11$) and the second stage (estimated marginal $M = 6.04$, $SE = 0.13$).

The interaction effect between the learning stage and condition was not significant,

$F(1, 95) = 0.00$, $p = .986$, $\eta^2 p = .00$, indicating that participants allocated to different conditions had similar trends of change in germane processing between the two learning stages.

8.3 Discussion

8.3.1 How Did Tracing Conditions Affect Participants' Learning Performance?

The analyses of the learning performance in Experiment 3 were more complicated than those in Experiment 1 and Experiment 2. The learning was divided into two stages in Experiment 3, resulting in two stages of practice performance. In the first stages, participants in both tracing conditions were requested to trace only via their right hand's index finger. The analysis of this stage was the same as Experiment 1 and Experiment 2; the pre-test score worked as the only covariate in the analysis. In the second stage, participants applied different tracing strategies according to their assigned condition. The pre-test score and the first-stage practice time were both included as covariates in the analysis of immediate and delayed post-tests. More indicators were added to the immediate post-test questions and the delayed post-test questions. Besides the post-test score and the post-test time that were previously applied in Experiment 1 and Experiment 2, analyses of similar test item scores and transfer test item scores were also conducted in Experiment 3.

Table 22 summarises participants' accuracy rates at different stages in Experiment 3. Overall, the participants' accuracy rates exhibited an "N" shaped curve across the three phases. Participants in both groups had medium levels of accuracy in the first stage of the practice phase and then had the highest accuracy in the second stage of the practice phase, followed by lower accuracy in the immediate post-test phase, and a bit higher accuracy in the delayed post-test phase. These findings were a little different from the trend we found in Experiment 1 and Experiment 2, where participants had the highest accuracy rate in the delayed post-test. In all three experiments, better performance in the delayed post-test suggested the possible presence of a practice effect (Anastasi, 1988; Campbell & Stanley, 1963). However, in Experiment 3, the specific questions in the immediate post-test and delayed post-test were no longer identical, which may cause the practice effect not to be as strong as in Experiment 1 and Experiment 2. This also explained why participants' delayed post-test accuracy rate in Experiment 3 was lower than in Experiment 1 and Experiment 2. Further repeated measure analyses suggested that participants performed better in the delayed post-test when measuring overall test score, overall test time to solution, and similar question score. However, there was no significant difference in participants' transfer question scores between the immediate post-test and the delayed post-test.

Table 22

Participants' Accuracy Rates in Three Tests in Experiment 3

Practice Questions			
	Total	Non-Alternating	Alternating
First Stage Accuracy	79.03%	78.47%	79.60%
Second Stage Accuracy	92.43%	93.77%	91.17%
Immediate Post-Test Questions			
	Total	Non-Alternating	Alternating
Similar Questions Accuracy	58.93%	57.47%	60.38%
Transfer Questions Accuracy	56.31%	53.39%	59.19%
General Accuracy	57.89%	55.84%	59.90%
Delayed Post-Test Questions			
	Total	Non-Alternating	Alternating
Similar Questions Accuracy	73.44%	70.63%	76.25%
Transfer Questions Accuracy	58.44%	55.63%	61.25%
General Accuracy	65.94%	63.13%	68.75%

In group comparisons, previous analysis showed that participants in alternating tracing conditions had similar learning performances in most tests after considering their prior calculation and learning abilities. The only detected difference was participants' transfer question accuracy rate in the immediate post-test and participants' delayed post-test score, in which participants in the alternating condition performed marginally better than those in the non-alternating condition. Though the effect was weak, it supported our assumption that alternating tracing could further support learning. Participants in the alternating condition had a marginally better performance in the delayed post-test than those in the non-alternating condition, supporting the speculation in previous experiments that the alternating fingers tracing strategy took longer to manifest its effect. Participants in the alternating condition had a marginally better performance on transfer questions in the immediate post-test than those in the non-alternating condition, suggesting that alternating fingers tracing may promote learning at a deeper level.

8.3.2 How Did Tracing Conditions Affect Participants' Intrinsic Motivation and Cognitive Load?

Previous analysis described participants' intrinsic motivation and three types of cognitive load in two different learning stages. Repeated-measure analyses revealed the change in motivation and cognitive load between stages.

Generally, participants reported significantly lower intrinsic motivation in the second stage. First-stage learning material and second-stage learning material were designed in a similar structure, which explained the decline in intrinsic motivation among participants after viewing similar learning content. The interaction effect was also found between stage and condition, suggesting that participants in the non-alternating condition had a greater decline in their intrinsic motivation during the two learning stages than those in the alternating condition. This finding, to some extent, aligned with the initial hypothesis that alternating finger tracing may trigger higher intrinsic motivation. This finding also suggested that non-alternating tracing and alternating tracing strategies could be included in the instructional design together to maintain learners' intrinsic motivation.

Participants also reported a significantly lower intrinsic cognitive load in the second stage. The interaction effect between the learning stage and participants' learning ability was found, suggesting that participants with lower learning ability have a slight decline in their intrinsic cognitive load during the two learning stages. This finding was consistent with the basic tenet of cognitive load theory that the intrinsic cognitive load was related to learners' prior ability (Sweller et al., 2011).

Participants' extraneous cognitive load showed no general difference between the two learning stages, but the interaction effect between the learning stage and condition was marginally significant. Participants in the non-alternating condition have lower extraneous cognitive load in the second stage compared to the first stage, while participants in the alternating condition have higher extraneous cognitive load in the second stage compared to the first stage. This finding suggested that continuous exposure to similar learning materials may decrease learners' extraneous cognitive load because learners can better ignore the irrelevant burden.

No significant difference was found between participants' germane processing in the two stages. The interaction effect was also not significant. This finding suggested that participants in the two conditions possessed cognitive learning at the same level.

The goal of the present study is to compare the differences between non-alternating tracing and alternating tracing, discovering the alternating finger tracing's further effect on supporting cognitive learning. For this reason, participants were instructed to make a set of number of tracing actions at each step in the worked examples in every experiment. Across Experiment 1 to Experiment 3, video analyses were used to assess the extent to which participants complied with the instructions. However, these analyses coincidentally revealed substantial variation in the number of tracing actions actually made by participants. In the next chapter, this data was explored to investigate potential dosage effects.

Chapter Nine: Supplementary Analysis on Tracing Dosage

In all three experiments, the tracing instruction required participants to trace five times for each specified item in the learning material, which was designed to standardise the dosage of tracing. However, video coding based on the coding scheme (see Appendix B) identified considerable variation in participants' tracing behaviours. Though all participants were given two minutes to learn each worked example, and all of them received the same instruction about tracing five times, their different learning speeds and habits resulted in varying numbers of tracing actions. Thus, there was an opportunity here to explore the relations between the dosage of tracing actions and subsequent learning performance.

Dose-response relationships have always been a key indicator in medical and clinical fields. Dose-response research is common in pharmaceutical contexts (Phillips, 1997), and other medical interventions have also been guided by dosage research (Payne et al., 2025). Correspondingly, dosage research is also meaningful in the field of education, giving teachers and educators clear guidance on how to apply an instructional method. Itoh et al. (2022) examined how many imagery sessions per week were most beneficial for basketball players to improve their free-throw shooting performance, finding that four imagery sessions per week, instead of five imagery sessions per week (the biggest dosage in the experiment), was most helpful. This experiment suggested the dose-response protocol had potential application in the educational field. In terms of CLT-based research, some tracing studies also intentionally or unintentionally formed groups with different dosages of tracing. For example, Hu et al. (2015) formed three conditions (no tracing, tracing above paper, and tracing on paper) in the study, which can be regarded as giving orderly increased treatment levels. Wang et al. (2022) had a similar design on ordered dosage levels, using three conditions (no tracing, tracing/tracing, and tracing/imagination) in the experiment. These studies hint at the potential importance of tracing's dosage. Thus, the current research made the following explanatory analyses to further investigate how the dosage of tracing affects cognitive learning performance.

In contrast with previous experiments focusing on group differences, dosage analyses discussed in this chapter aim to explore the potential relationship between tracing dosage and learning performance. Thus, a new research question - *How does variation in tracing actions affect learning performance?* - is addressed in this exploratory research. Those participants excluded in the above analyses for Experiments 1-3 due to non-compliance were included in the dosage analyses because these less-than-compliant tracing actions reflect the natural variation in tracing actions seen among participants. Those participants who did not fulfil the group condition (e.g., applying non-alternating tracing in the alternating condition) were still excluded from the analyses because these data may interfere with the separate analyses within each group.

Six indicators were coded to measure participants' dosage (see Appendix B), including two indicators (Category C: Tracing Counts, and D: Total Tracing) to measure participants' tracing dosage across both groups; two indicators (Category A: Dominant Tracing, and F: Finger Continued Counts) to measure participants' tracing dosage separately in both groups; and two indicators (Category B: Non-Dominant Tracing, and E: Finger Alternated Counts) to measure tracing dosage in the alternating groups only.

The dosage analyses of three experiments are described in the following sections.

9.1 Experiment 1

One hundred and fifteen participants joined and completed the experiment. Four participants were excluded from the dosage analyses because of incorrect tracing action (numbers 43, 48, 68, 105), and one participant was removed because of data corruption (number 46, 30-second video recording was lost in worked example 4). The following data analysis was based on the remaining 110 participants.

9.1.1 General Tracing and Learning Performance

The general analyses on tracing dosage and learning performance were based on participants in both conditions (110 participants). Two indicators, the number of times the participants learned and traced the entire worked example (Tracing Counts) and the total number of elliptical tracing actions of both hands made (Total Tracing), were analysed. Though these two indicators were highly correlated with each other ($r = .89$), it was necessary to include both of them because they represented two different tracing actions. For example, one participant learned the worked example four times and traced each figure five times as instructed in the two-minute learning; the tracing count was four, and the total tracing was 80 for this participant. However, another participant learned the worked example four times but traced each figure 20 times; the tracing count was one, and the total tracing was also 80 for this participant. They are two completely different tracing actions and cannot be reflected solely through the key indicator total tracing. Thus, the tracing count was necessary to be kept in the analyses.

As participants' pre-test scores were significantly correlated with their learning performance, pre-test performance need to be controlled for when considering dosage effects on learning performance. Thus, partial correlational analysis was conducted to assess whether variation in tracing actions correlated with participants' learning performance, controlling for participants' pre-test score. As Table 23 shows, participants' tracing counts and total tracing were significantly correlated with participants' learning performance in the practice phase, immediate post-test phase, and the delayed post-test phase.

Table 23

Partial Correlations Between Both Group Participants' Tracing Counts, Total Tracing, and Learning Performance in Experiment 1

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8
1. Tracing Counts	110	8.72	4.67	—							
2. Total Tracing	110	161.03	96.12	.88***	—						
3. Practice Score (/4)	110	2.67	1.24	.24*	.23*	—					
4. Practice Time to Solution (/120)	110	89.56	21.63	-.32***	-.29**	-.81***	—				
5. Immediate Post-Test Score (/20)	110	11.88	5.58	.22*	.24*	.48***	-.37***	—			
6. Immediate Post-Test Time to Solution (/400)	110	348.14	38.24	-.23*	-.21*	-.43***	.49***	-.77***	—		
7. Delayed Post-Test Score (/20)	110	15.97	4.03	.21*	.19*	.41***	-.26***	.59***	-.37***	—	
8. Delayed Post-Test Time to Solution (/400)	110	308.22	44.40	-.25**	-.20*	-.39***	.44***	-.64***	.78***	-.64***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analyses were also conducted for participants' intrinsic motivation and cognitive load self-reports, controlling for pre-test scores. Results show that tracing counts and total tracing did not correlate with participants' intrinsic motivation or cognitive load (see Table 24).

In summary, participants' tracing counts and total tracing times significantly predicted their learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase, but did not correlate with participants' intrinsic motivation and cognitive load.

Table 24

Partial Correlations Between Both Group Participants' Tracing Counts, Total Tracing, Intrinsic Motivation, and Cognitive Load in Experiment 1

Variable	*n*	*M*	*SD*	1	2	3	4	5	6
1. Tracing Counts	110	8.72	4.67	—					
2. Total Tracing	110	161.03	96.12	.88***	—				
3. Intrinsic Motivation (/8)	110	5.73	1.53	-.04	-.13	—			
4. Intrinsic Cognitive Load (/8)	110	1.20	1.29	-.14	-.11	-.12	—		
5. Extraneous Cognitive Load (/8)	110	1.80	1.49	-.02	.02	-.10	.67***	—	
6. Germane Processing (/8)	110	6.22	1.19	.12	.07	.27**	-.57***	-.61***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

9.1.2 Non-Alternating Finger Tracing and Learning Performance

Since the participants were randomly assigned to two experimental conditions and applied different tracing strategies, the dosage analyses can also be further analysed within each condition. Analyses for the non-alternating condition (55 participants) focus on three indicators: total number of dominant hands' elliptical tracing made (Dominant Tracing), the number of times the participants learned and traced the entire worked example (Tracing Counts) and the number of times the participants paused or changed which figure was being traced during the same index finger tracing (Finger Continued Counts). As participants in the non-alternating condition only used their dominant hand to trace, the total number of dominant-hand elliptical tracing made was equal to the total number of elliptical tracing actions made by the participants.

As Table 25 shows, dominant tracing was significantly correlated with all learning performance indicators. However, participants' tracing counts and finger continued counts were not correlated with the practice score. Considering the indicator test time to solution had a more differentiated distribution, it was reasonable that some factors only correlate with the test time to solution, but not the test score. When there was no significant correlation found in test score, findings on test time to solution also represented the correlation between participants' learning performance and the corresponding variable.

Table 25

Partial Correlations Between Participants' Dominant Tracing, Tracing Counts, Finger Continued Counts, and Learning Performance in the Non-Alternating Group in Experiment 1

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Dominant Tracing	55	160.83	99.79	—								
2. Tracing Counts	55	9.20	5.12	.92***	—							
3. Finger Continued Counts	55	35.73	23.03	.83***	.94***	—						
4. Practice Score (/4)	55	2.82	1.23	.29*	.22	.22	—					
5. Practice Time to Solution (/120)	55	87.05	21.20	-.35*	-.30*	-.27*	-.82***	—				
6. Immediate Post-Test Score (/20)	55	12.02	5.64	.39**	.36*	.29*	.40**	-.37**	—			
7. Immediate Post-Test Time to Solution (/400)	55	346.12	39.43	-.36**	-.36*	-.27*	-.29**	.38**	-.72***	—		
8. Delayed Post-Test Score (/20)	55	15.80	3.95	.28*	.31*	.28*	.45***	-.31*	.63***	-.36**	—	
9. Delayed Post-Test Time to Solution (/400)	55	307.37	43.25	-.39**	-.43**	-.36**	-.32*	.40**	-.62***	.82***	-.60***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analysis was also conducted on participants' intrinsic motivation and cognitive load, suggesting that participants' dominant tracing, tracing counts, and finger continued counts did not correlate with participants' intrinsic motivation and cognitive load (see Table 26).

In summary, participants' tracing counts and total tracing times were significantly correlated with their learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase, but did not correlate with participants' intrinsic motivation and cognitive load.

Table 26

Partial Correlation Between Participants' Dominant Tracing, Tracing Counts, Finger Continued Counts, Intrinsic Motivation, and Cognitive Load in the Non-Alternating Group in Experiment 1

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7
1. Dominant Tracing	55	160.83	99.79	—						
2. Tracing Counts	55	9.20	5.12	.92***	—					
3. Finger Continued Counts	55	35.73	23.03	.83***	.94***	—				
4. Intrinsic Motivation (/8)	5.68	1.64	5.68	-.14	-.01	-.02	—			
5. Intrinsic Cognitive Load (/8)	1.04	1.37	1.04	-.16	-.13	-.09	-.14	—		
6. Extraneous Cognitive Load (/8)	1.59	1.51	1.59	.05	.09	.09	-.16	.72***	—	
7. Germane Processing (/8)	6.33	1.08	6.33	-.02	.01	.01	.45***	-.52***	-.52***	—

* *p* < .05, ** *p* < .01, *** *p* < .001

9.1.3 Alternating Finger Tracing and Learning Performance

Tracing dosage effect analyses for the alternating fingers tracing strategy were based on participants in the alternating condition (55 participants). All six indictors, the number of dominant hands' elliptical tracing made (Dominant Tracing), the number of non-dominant hands' elliptical tracing made (Non-Dominant Tracing), the number of times the participants learned and traced the entire worked example (Tracing Counts), the total number of elliptical tracing actions of both hands made (Total Tracing), the number of times the participants switched between their two hands' index fingers to trace (Finger Switched Counts), and the number of times the participants paused or changed the figures during the same index finger tracing (Finger Continued Counts), were analysed.

A partial correlational analysis was conducted to explore to further check if tracing dosage correlated with participants' learning performance in the alternating tracing group. As Table 27 shows, among six tracing dosage factors, only tracing counts and finger alternated counts correlated with participants' test time to solution in the practice phase.

Table 27

Partial Correlations Between Six Tracing Dosage Factors and Learning Performance in the Alternating Group in Experiment 1

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. Dominant Tracing	55	80.60	45.70	—											
2. Non-dominant Tracing	55	80.62	48.72	.95***	—										
3. Tracing Counts	55	8.24	4.17	.86***	.82***	—									
4. Total Tracing	55	161.21	93.25	.99***	.99***	.85***	—								
5. Finger Alternated Counts	55	16.31	8.52	.88***	.84***	.98***	.87***	—							
6. Finger Continued Counts	55	14.22	8.69	.77***	.73***	.89***	.76***	.85***	—						
7. Practice Score (/4)	55	2.53	1.25	.17	.17	.24	.17	.21	.11	—					
8. Practice Time to Solution (/120)	55	92.06	21.96	-.26	-.26	-.34*	-.27	-.31*	-.27	-.78***	—				
9. Immediate Post-Test Score (/20)	55	11.75	5.56	.07	.08	.01	.08	.02	-.04	.52***	-.33*	—			
10. Immediate Post-Test Time to Solution (/400)	55	350.17	37.27	-.03	-.07	-.04	-.05	-.02	-.06	-.51***	.53***	-.80***	—		
11. Delayed Post-Test Score (/20)	55	16.15	4.15	.12	.14	.11	.13	.10	.00	.38**	-.22	.57***	-.37***	—	
12. Delayed Post-Test Time to Solution (/400)	55	309.08	45.90	-.01	-.03	-.03	-.02	-.04	-.01	-.42**	.44***	-.64***	.75***	-.68***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analysis was also conducted on participants' intrinsic motivation and cognitive load, suggesting that all six tracing factors did not correlate with participants' intrinsic motivation and cognitive load (see Table 28).

In summary, potential correlation may exist between tracing dosage and alternating group participants' learning performance in the practice phase. The only significant correlation between tracing dosage and learning performance was reflected by tracing counts and finger alternated counts, indicating that compared to how many times of elliptical tracing was made, how many times participants traced the entire worked example and alternated their hands were more important. When considering tracing's dosage effect, more repetitions across the worked example may prove more valuable than merely repeating the tracing actions. However, the correlation was not observed in the immediate post-test phase and the delayed post-test phase. No correlation was found on participants' intrinsic motivation and cognitive load.

Table 28

Partial Correlations Between Six Tracing Dosage Factors, Intrinsic Motivation, and Cognitive Load in the Alternating Group in Experiment 1

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Dominant Tracing	55	80.60	45.70	—									
2. Non-dominant Tracing	55	80.62	48.72	.95***	—								
3. Tracing Counts	55	8.24	4.17	.86***	.82***	—							
4. Total Tracing	55	161.21	93.25	.99***	.99***	.85***	—						
5. Finger Alternated Counts	55	16.31	8.52	.88***	.84***	.98***	.87***	—					
6. Finger Continued Counts	55	14.22	8.69	.77***	.73***	.89***	.76***	.85***	—				
7. Intrinsic Motivation (/8)	55	5.77	1.41	-.17	-.25	-.11	-.21	-.11	-.11	—			
8. Intrinsic Cognitive Load (/8)	55	1.36	1.20	-.05	.04	-.08	.00	-.11	.03	-.07	—		
9. Extraneous Cognitive Load (/8)	55	2.00	1.46	.01	.06	-.08	.03	-.09	-.01	.01	.57***	—	
10. Germane Processing (/8)	55	6.11	1.29	.16	.09	.22	.13	.25	.13	.09	-.62***	-.68***	—

* *p* < .05, ** *p* < .01, *** *p* < .00

9.2 Experiment 2

One hundred and twelve participants joined and completed the experiment. Three were excluded from the dosage analyses because of incorrect tracing action (numbers 52, 70, 72). The following data analysis was based on the remaining 109 participants.

9.2.1 General Tracing and Learning Performance

The general analyses on tracing dosage and learning performance were based on participants in both conditions (109 participants). Two indicators, the number of times the participants learned and traced the entire worked example (Tracing Counts) and the total number of elliptical tracing actions of both hands made (Total Tracing), were analysed.

A partial correlational analysis was conducted to explore to further check if tracing dosage correlated with participants' learning performance. As Table 29 shows, participants' tracing counts and total tracing were significantly correlated with participants' practice score, practice time to solution, immediate post-test score, immediate post-test to solution, and delayed post-test score, but did not correlate with the delayed post-test time to solution.

Table 29

Partial Correlations Between Both Group Participants' Tracing Counts, Total Tracing, and Learning Performance in Experiment 2

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8
1. Tracing Counts	109	9.79	5.80	—							
2. Total Tracing	109	198.23	125.98	.92***	—						
3. Practice Score (/4)	109	2.81	1.13	.29**	.30**	—					
4. Practice Time to Solution (/120)	109	91.97	17.29	-.32***	-.35***	-.78***	—				
5. Immediate Post-Test Score (/20)	109	11.41	5.04	.27**	.28**	.41***	-.47***	—			
6. Immediate Post-Test Time to Solution (/400)	109	349.08	33.06	-.29**	-.31**	-.40***	.51***	-.84***	—		
7. Delayed Post-Test Score (/20)	109	15.57	5.07	.20*	.20*	.39***	-.36***	.62***	-.46***	—	
8. Delayed Post-Test Time to Solution (/400)	109	302.54	51.27	-.12	-.13	-.36***	.49***	-.67***	.69***	-.43***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analysis was also conducted on participants' intrinsic motivation and cognitive load, suggesting that tracing counts and total tracing did not correlate with participants' intrinsic motivation and cognitive load (see Table 30).

In summary, participants' tracing counts and total tracing times were significantly correlated with learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase, but did not correlate with participants' intrinsic motivation or cognitive load.

Table 30

Partial Correlations Between Both Group Participants' Tracing Counts, Total Tracing, Intrinsic Motivation, and Cognitive Load in Experiment 2

Variable	*n*	*M*	*SD*	1	2	3	4	5	6
1. Tracing Counts	109	9.79	5.80	—					
2. Total Tracing	109	198.23	125.98	.92***	—				
3. Intrinsic Motivation (/8)	109	5.59	1.65	-.09	-.06	—			
4. Intrinsic Cognitive Load (/8)	109	1.99	1.80	-.16	-.17	-.41***	—		
5. Extraneous Cognitive Load (/8)	109	2.57	1.68	-.06	-.06	-.43***	.73***	—	
6. Germane Processing (/8)	109	5.67	1.62	.06	.09	.60***	-.68***	-.61***	—

* *p* < .05, ** *p* < .01, *** *p* < .001

9.2.2 Non-Alternating Finger Tracing and Learning Performance

Since the participants were allocated into two conditions and applied different tracing strategies, the dosage analyses can also be further analysed within each condition. The analyses about tracing dosage effect when participants applied the non-alternating finger tracing strategy were based on participants in the non-alternating condition (56 participants). Three indicators, total number of dominant hands' elliptical tracing made (Dominant Tracing), the number of times the participants learned and traced the entire worked example (Tracing Counts) and the number of times the participants paused or changed the figures during the same index finger tracing (Finger Continued Counts), were analysed.

As Table 31 shows, participants' finger continued counts was significantly correlated with participants' test score and test time to solution in the practice phase.

However, participants' dominant tracing was only correlated with the practice test time to solution while tracing counts was only correlated with the practice score. No other correlation was found between the three indicators and participants' learning performance in the immediate post-test and delayed post-test.

Table 31

Partial Correlation Between Participants' Dominant Tracing, Tracing Counts, Finger Continued Counts, and Learning Performance in the Non-Alternating Group in Experiment 2

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Dominant Tracing	56	189.19	128.51	—								
2. Tracing Counts	56	9.37	5.63	.89***	—							
3. Finger Continued Counts	56	40.21	26.26	.77***	.88***	—						
4. Practice Score (/4)	56	11.77	4.84	.26	.28*	.29*	—					
5. Practice Time to Solution (/120)	56	349.40	32.16	-.29*	-.23	-.29*	-.79***	—				
6. Immediate Post-Test Score (/20)	56	11.77	4.84	.24	.19	.15	.23	-.35**	—			
7. Immediate Post-Test Time to Solution (/400)	56	349.40	32.16	-.09	-.06	-.12	-.22	.36**	-.81***	—		
8. Delayed Post-Test Score (/20)	56	16.05	4.59	.12	.15	.13	.34*	-.37**	.63***	-.43***	—	
9. Delayed Post-Test Time to Solution (/400)	56	304.48	39.73	-.07	-.07	-.11	-.20	.34*	-.66***	.69***	-.78***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analysis was also conducted on participants' intrinsic motivation and cognitive load, suggesting that participants' dominant tracing, tracing counts, and finger continued counts did not correlate with participants' intrinsic motivation and cognitive load (see Table 32).

In summary, participants' dominant tracing, tracing counts, and finger continued counts were significantly correlated with the learning performance in the practice phase. Among the three indicators, only the dominant tracing variable was significantly correlated with learning performance in the immediate post-test phase. None of the three indicators was correlated with the learning performance in the delayed post-test phase. None of the three indicators correlated with participants' intrinsic motivation or cognitive load.

Table 32

Partial Correlations Between Participants' Dominant Tracing, Tracing Counts, Finger Continued Counts, Intrinsic Motivation, and Cognitive Load in the Non-Alternating Group in Experiment 2

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7
1. Dominant Tracing	56	189.19	128.51	—						
2. Tracing Counts	56	9.37	5.63	.89***	—					
3. Finger Continued Counts	56	40.21	26.26	.77***	.88***	—				
4. Intrinsic Motivation (/8)	56	5.76	1.66	-.03	-.07	-.22	—			
5. Intrinsic Cognitive Load (/8)	56	1.83	1.78	-.20	-.22	-.08	-.23	—		
6. Extraneous Cognitive Load (/8)	56	2.28	1.62	-.07	-.10	-.04	-.25	.68***	—	
7. Germane Processing (/8)	56	5.81	1.50	.10	.06	-.11	.57***	-.56***	-.50***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

9.2.3 Alternating Finger Tracing and Learning Performance

The analyses of tracing dosage when participants applied the alternating fingers tracing strategy were based on participants in the alternating condition (53 participants). All six indictors, the number of dominant hands' elliptical tracing made (Dominant Tracing), the number of non-dominant hands' elliptical tracing made (Non-Dominant Tracing), the number of times the participants learned and traced the entire worked example (Tracing Counts), the total number of elliptical tracing actions of both hands made (Total Tracing), the number of times the participants switched between their two hands' index fingers to trace (Finger Switched Counts), and the number of times the participants paused or changed the figures during the same index finger tracing (Finger Continued Counts), were analysed.

As Table 33 shows, all six tracing dosage factors were significantly correlated with the test score and test time to solution in the practice phase. In the immediate post-test phase, dominant tracing, non-dominant tracing, tracing counts, total tracing, and finger alternated counts were correlated with participants' immediate post-test scores and immediate post-test time to solution, but only finger continued counts was correlated with participants' immediate post-test time to solution. No significant correlations were found between the six factors and participants' test scores or test time to solution in the delayed post-test phase.

Table 33

Partial Correlations Between Six Tracing Dosage Factors and Learning Performance in the Alternating Group in Experiment 2

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. Dominant Tracing	53	105.21	64.21	—											
2. Non-dominant Tracing	53	102.57	60.96	.96***	—										
3. Tracing Counts	53	10.24	6.00	.96***	.93***	—									
4. Total Tracing	53	207.78	123.75	.99***	.99***	.96***	—								
5. Finger Alternated Counts	53	20.72	13.26	.97***	.96***	.98***	.97***	—							
6. Finger Continued Counts	53	21.77	16.65	.83***	.85***	.75***	.85***	.79***	—						
7. Practice Score (/4)	53	2.79	1.08	.30*	.31*	.29*	.30*	.28*	.30*	—					
8. Practice Time to Solution (/120)	53	92.45	16.18	-.42**	-.43**	-.40**	-.43**	-.40**	-.45***	-.79***	—				
9. Immediate Post-Test Score (/20)	53	11.04	5.26	.31*	.28*	.31*	.30*	.28*	0.22	.53***	-.60***	—			
10. Immediate Post-Test Time to Solution (/400)	53	349.70	34.71	-.48***	-.44***	-.46***	-.47***	-.45***	-.36**	-.52***	.67***	-.85***	—		
11. Delayed Post-Test Score (/20)	53	15.06	5.52	0.25	0.24	0.24	0.25	0.23	0.24	.41**	-.35*	.60***	-.46**	—	
12. Delayed Post-Test Time to Solution (/400)	53	300.48	61.49	-0.15	-0.12	-0.13	-0.14	-0.14	-0.11	-.47***	.63***	-.69***	.70***	-.28*	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Partial correlational analysis was also conducted on participants' intrinsic motivation and cognitive load self-reports, indicating that the six tracing factors did not correlate with participants' intrinsic motivation and cognitive load (see Table 34).

In summary, significant correlations were found between tracing dosage and alternating group participants' learning performance in the practice phase and the immediate post-test phase. However, the correlation was not observed in the delayed post-test phase. No correlations were found for participants' intrinsic motivation and cognitive load.

Table 34

Partial Correlation Between Six Tracing Dosage Factors, Intrinsic Motivation, and Cognitive Load in the Alternating Group in Experiment 2

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Dominant Tracing	53	105.21	64.21	—									
2. Non-dominant Tracing	53	102.57	60.96	.96***	—								
3. Tracing Counts	53	10.24	6.00	.96***	.93***	—							
4. Total Tracing	53	207.78	123.75	.99***	.99***	.96***	—						
5. Finger Alternated Counts	53	20.72	13.26	.97***	.96***	.98***	.97***	—					
6. Finger Continued Counts	53	21.77	16.65	.83***	.85***	.75***	.85***	.79***	—				
7. Intrinsic Motivation (/8)	53	5.41	1.63	-.06	-.09	-.10	-.07	-.06	-.17	—			
8. Intrinsic Cognitive Load (/8)	53	2.16	1.83	-.13	-.09	-.08	-.11	-.11	.02	-.60***	—		
9. Extraneous Cognitive Load (/8)	53	2.90	1.69	-.05	-.02	-.02	-.04	-.04	.06	-.59***	.77***	—	
10. Germane Processing (/8)	53	5.53	1.73	.08	.04	.05	.06	.06	-.11	.64***	-.78***	-.70***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

9.3 Experiment 3

One hundred and eight participants joined and completed the experiment. Two were excluded from the dosage analyses because of incorrect tracing action (participant number 1, 44). The following data analysis was based on the remaining 106 participants.

9.3.1 General Tracing and Learning Performance

The general analyses on tracing dosage and learning performance were based on participants in both conditions (106 participants). Two indicators, the number of times the participants learned and traced the entire worked example (Tracing Counts) and the total number of elliptical tracing actions of both hands made (Total Tracing), were analysed.

Since the learning phase was divided into two stages in Experiment 3, participants' learning performance in the two stages was separately analysed. A partial correlational analysis was conducted to explore whether tracing dosage correlated with participants' learning performance in the first stage. Participants traced the worked examples as they learned, so the first-stage tracing dosage indicators were included in the analysis. As Table 35 shows, participants' tracing counts and total tracing were not correlated with participants' practice learning performance in the first stage.

Similar partial correlational analysis was also conducted based on stage 2 practice phase, the immediate post-test phase, and the delayed post-test phase. Since participants answered the second-stage practice questions, the immediate post-test questions, and the delayed post-test questions after completing all tracing actions, the tracing dosage indicators that reflect the sum of the two stages were included in the analysis. As Table 36 shows, participants' tracing counts and total tracing were not correlated with participants' learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase.

Table 35

Partial Correlations Between Both Group Participants' First Stage Tracing Counts, Total Tracing, and Learning Performance in Practice Phase in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4
1. First-Stage Tracing Counts	106	7.79	4.45	—			
2. First-Stage Total Tracing	106	147.36	95.40	.82***	—		
3. First-Stage Practice Score (/3)	106	2.38	0.82	0.02	.02	—	
4. First-Stage Practice Time to Solution (/90)	106	56.86	16.77	-.06	-.04	-.72***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 36

Partial Correlations Between Both Group Participants' Tracing Counts, Total Tracing, and Learning Performance in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. Tracing Counts	106	16.9	9.7	—											
2. Total Tracing	106	313.0	193.2	.81***	—										
3. Second-Stage Practice Score (/3)	106	2.7	0.5	.04	.01	—									
4. Second-Stage Practice Time to Solution (/90)	106	44.9	12.0	.05	.14	-.49***	—								
5. Immediate Post-Test Score (/20)	106	11.2	5.0	.01	.01	.21*	-.25*	—							
6. Immediate Post-Test Similar Score (/12)	106	6.9	2.8	-.02	-.07	.21*	-.29**	.87***	—						
7. Immediate Post-Test Transfer Score (/8)	106	4.4	2.4	.05	.08	.21*	-.26**	.84***	.64***	—					
8. Immediate Post-Test Time to Solution (/400)	106	333.7	41.6	.01	-.01	-.17	.29**	-.75***	-.73***	-.75***	—				
9. Delayed Post-Test Score (/20)	105	12.9	4.9	.05	-.01	.21*	-.20*	.57***	.53***	.56***	-.56***	—			
10. Delayed Post-Test Similar Score (/10)	105	7.2	2.4	.06	-.04	.21*	-.16	.47***	.48***	.42***	-.34***	.88***	—		
11. Delayed Post-Test Transfer Score (/10)	105	5.7	2.9	.03	.02	.17	-.20*	.54***	.47***	.57***	-.47***	.90***	.59***	—	
12. Delayed Post-Test Time to Solution (/400)	105	310.4	51.5	-.12	-.08	-.15	.17	-.51***	-.45***	-.54***	.66***	-.73***	-.58***	-.72***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

A similar partial correlational analysis was also conducted on participants' first-stage intrinsic motivation and cognitive load. As participants answered the first-stage self-report survey after completing three worked examples, the first-stage tracing dosage indicators were included in the analysis. As Table 37 shows, first-stage tracing counts and total tracing did not correlate with participants' first-stage intrinsic motivation and cognitive load.

Table 37

Partial Correlations Between Both Group Participants' First-Stage Tracing Counts, Total Tracing, Intrinsic Motivation, and Cognitive Load in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6
1. First-Stage Tracing Counts	106	7.79	4.45	—					
2. First-Stage Total Tracing	106	147.36	95.40	.82***	—				
3. First-Stage Intrinsic Motivation (/8)	106	5.70	1.47	.07	.01	—			
4. First-Stage Intrinsic Cognitive Load (/8)	106	1.09	1.15	-.12	-.06	-.26**	—		
5. First-Stage Extraneous Cognitive Load (/8)	106	1.49	1.15	-.16	-.05	-.35***	.70***	—	
6. First-Stage Germane Processing (/8)	106	6.24	1.08	.01	-.11	.63***	-.52***	-.57***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

A similar partial correlational analysis was also conducted on participants' second-stage intrinsic motivation and cognitive load. As participants answered the second-stage self-report survey after completing all six worked examples, the tracing dosage indicators that reflect the sum of the two stages were included in the analysis. As Table 38 shows, first-stage tracing counts and total tracing did not correlate with participants' first-stage intrinsic motivation and cognitive load.

Table 38

Correlation Between Both Group Participants' Tracing Counts, Total Tracing, Second-Stage Intrinsic Motivation, and Cognitive Load in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6
1. Tracing Counts	106	16.88	9.70	—					
2. Total Tracing	106	312.99	193.17	.81***	—				
3. Second-Stage Intrinsic Motivation (/8)	106	5.21	1.74	.07	.00	—			
4. Second-Stage Intrinsic Cognitive Load (/8)	106	0.86	1.03	-.02	.04	-.07	—		
5. Second-Stage Extraneous Cognitive Load (/8)	106	1.51	1.35	-.04	.06	-.21*	.47***	—	
6. Second-Stage Germane Processing (/8)	106	6.06	1.20	.05	-.04	.52***	-.19	-.24*	—

* $p < .05$, ** $p < .01$, *** $p < .001$

In summary, participants' tracing counts and total tracing times did not correlate with learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase, and did not correlate with participants' intrinsic motivation and cognitive load.

9.3.2 Non-Alternating Finger Tracing and Learning Performance

Since the participants were allocated into two conditions and applied different tracing strategies, the dosage analyses can also be further analysed within each condition. The analyses about tracing dosage effect when participants applied the non-alternating finger tracing strategy were based on participants in the non-alternating condition (54 participants). Three indicators, total number of dominant hands' elliptical tracing made (Dominant Tracing), the number of times the participants learned and traced the entire worked example (Tracing Counts) and the number of times the participants paused or changed the figures during the same index finger tracing (Finger Continued Counts), were analysed.

A partial correlational analysis was conducted to further check if tracing dosage correlated with participants' learning performance in the non-alternating tracing group in the first-stage practice phase. The first-stage tracing dosage indicators were included in the analysis. As Table 39 shows, participants' first-stage dominant tracing, tracing counts, and finger continued counts were not correlated with participants' first-stage learning performance in the practice phase.

Table 39

Partial Correlations Between Participants' First Stage Tracing Counts, Total Tracing, and Learning Performance in Practice Phase in Non-Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5
1. First-Stage Dominant Tracing	54	157.37	111.26	—				
2. First-Stage Tracing Counts	54	7.98	4.56	.78***	—			
3. First-Stage Continued Counts	54	32.07	18.91	.76***	.95***	—		
4. First-Stage Practice Score (/3)	54	2.37	0.81	-.08	.03	.02	—	
5. First-Stage Practice Time to Solution (/90)	54	56.21	15.89	-.09	-.17	-.17	-.71***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analysis was also conducted on the second practice phase, the immediate post-test phase, and the delayed post-test phase. The tracing dosage indicators that reflect the sum of the two stages were included in the analysis. As Table 40 shows, participants' dominant tracing, tracing counts, and finger continued counts were not correlated with participants' learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase.

Table 40

Partial Correlation Between Participants' Tracing Counts, Total Tracing, and Learning Performance in Non-Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	
1. Dominant Tracing	54	334.09	230.29	—												
2. Tracing Counts	54	17.74	10.60	.79***	—											
3. Finger Continued Counts	54	69.43	40.80	.79***	.96***	—										
4. Second-Stage Practice Score (/3)	54	2.74	0.48	-.02	.06	.08	—									
5. Second-Stage Practice Time to Solution (/90)	54	45.20	12.87	.19	.05	.05	-.52***	—								
6. Immediate Post-Test Score (/20)	54	10.80	5.28	.03	.08	.11	.21	-.31*	—							
7. Immediate Post-Test Similar Score (/12)	54	6.70	3.12	-.10	.01	-.04	.15	-.31*	.91***	—						
8. Immediate Post-Test Transfer Score (/8)	54	4.09	2.48	.17	.14	.17	.24	-.24	.88***	.61***	—					
9. Immediate Post-Test Time to Solution (/400)	54	336.52	41.45	-.08	-.07	-.12	-.18	.37**	-.87***	-.77***	-.80***	—				
10. Delayed Post-Test Score (/20)	54	12.24	5.46	.01	.09	.12	.28*	-.21	.63***	.60***	.53***	-.51***	—			
11. Delayed Post-Test Similar Score (/10)	54	6.85	2.74	-.05	.07	.09	.26	-.16	.54***	.54***	.42***	-.43**	.91***	—		
12. Delayed Post-Test Transfer Score (/10)	54	5.39	3.11	.07	.08	.12	.25	-.22	.61***	.54***	.55***	-.51***	.91***	.66***	—	
13. Delayed Post-Test Time to Solution (/400)	54	312.13	52.51	-.07	-.10	-.13	-.23	.23	-.64***	-.57***	-.58***	.69***	-.83***	-.69***	-.82***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analyses were also conducted on participants' first-stage and second-stage intrinsic motivation and cognitive load. The first-stage tracing dosage indicators were included in the analysis. As Table 41 and Table 42 show, first-stage and second-stage dominant tracing, tracing counts, and finger continued counts did not correlate with participants' first-stage or second-stage intrinsic motivation and cognitive load.

Table 41

Partial Correlations Between Participants' First-Stage Dominant Tracing, Tracing Counts, Finter Continued Counts, Intrinsic Motivation, and Cognitive Load in Non-Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	
1. First-Stage Dominant Tracing	54	157.37	111.26	—						
2. First-Stage Tracing Counts	54	7.98	4.56	.78***	—					
3. First-Stage Finger Continued Counts	54	32.07	18.91	.76***	.95***	—				
4. First-Stage Intrinsic Motivation (/8)	54	5.80	1.44	.04	.14	.06	—			
5. First-Stage Intrinsic Cognitive Load (/8)	54	0.93	0.96	.04	-.02	-.02	-.10	—		
6. First-Stage Extraneous Cognitive Load (/8)	54	1.49	1.11	.00	-.19	-.18	-.27*	.64***	—	
7. First-Stage Germane Processing (/8)	54	6.24	1.13	-.19	-.01	-.06	.67***	-.36**	-.49***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 42

Partial Correlations Between Participants' Dominant Tracing, Tracing Counts, Finger Continued Counts, Second-Stage Intrinsic Motivation, and Cognitive Load in the Non-Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7
1. Dominant Tracing	54	334.09	230.29	—						
2. Tracing Counts	54	17.74	10.60	.79***	—					
3. Finger Continued Counts	54	69.43	40.80	.79***	.96***	—				
4. Intrinsic Motivation (/8)	54	5.08	1.82	.07	.18	.18	—			
5. Intrinsic Cognitive Load (/8)	54	0.78	0.98	.08	-.01	.04	-.12	—		
6. Extraneous Cognitive Load (/8)	54	1.31	1.30	.08	-.09	-.11	-.13	.62***	—	
7. Germane Processing (/8)	54	6.07	1.31	-.04	.10	.08	.70***	-.07	-.09	—

* $p < .05$, ** $p < .01$, *** $p < .001$

9.3.3 Alternating Finger Tracing and Learning Performance

The analyses for tracing dosage effects when participants applied the alternating fingers tracing strategy were based on participants in the alternating condition (53 participants). All six indictors, the number of dominant hands' elliptical tracing made (Dominant Tracing), the number of non-dominant hands' elliptical tracing made (Non-Dominant Tracing), the number of times the participants learned and traced the entire worked example (Tracing Counts), the total number of elliptical tracing actions of both hands made (Total Tracing), the number of times the participants switched between their two hands' index fingers to trace (Finger Switched Counts), and the number of times the participants paused or changed the figures during the same index finger tracing (Finger Continued Counts), were analysed.

A partial correlational analysis was conducted to explore to further check if tracing dosage correlated with participants' learning performance in the alternating tracing group in the first-stage practice phase. As participants in the alternating condition also applied the non-alternating tracing strategy in the first-stage learning, only dominant tracing, tracing counts, and finger continued counts were included in the analysis. Similar to previous analysis, the first-stage tracing dosage indicators were included in the analysis. As Table 43 shows, participants' first-stage dominant tracing, tracing counts, and finger continued counts were not correlated with participants' first-stage learning performance in the practice phase.

Table 43

Partial Correlations Between Participants' First Stage Tracing Counts, Total Tracing, and Learning Performance in Practice Phase in Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5
1. First-Stage Dominant Tracing	52	134.86	76.34	—				
2. First-Stage Tracing Counts	52	7.59	4.36	.91***	—			
3. First-Stage Continued Counts	52	28.87	17.32	.93***	.97***	—		
4. First-Stage Practice Score (/3)	52	2.39	0.84	.04	.02	.07	—	
5. First-Stage Practice Time to Solution (/90)	52	57.54	17.77	.06	.05	.00	-.73***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Similar partial correlational analysis was also conducted on the full practice phase, the immediate post-test phase, and the delayed post-test phase. All six tracing dosage indicators that reflect the sum of the two stages were included in the analysis. As Table 44 shows, all six tracing dosage indicators were not correlated with participants' learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase.

Table 44

Partial Correlations Between All Six Tracing Dosage Factors and Learning Performance in the Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Dominant Tracing	52	217.07	116.57	—															
2. Non-dominant Tracing	52	69.33	37.10	.81***	—														
3. Tracing Counts	52	15.99	8.68	.87***	.74***	—													
4. Total Tracing	52	286.40	148.18	.99***	.89***	.87***	—												
5. Finger Alternated Counts	52	16.60	9.70	.84***	.83***	.92***	.87***	—											
6. Finger Continued Counts	52	45.71	25.89	.88***	.73***	.96***	.87***	.90***	—										
7. Second-Stage Practice Score (/3)	52	2.71	0.46	.00	-.12	.01	-.03	-.13	.05	—									
8. Second-Stage Practice Time to Solution (/90)	52	44.56	11.22	.02	.04	.02	.02	.08	-.04	-.46***	—								
9. Immediate Post-Test Score (/20)	52	11.54	4.66	.06	-.07	-.02	.03	-.04	.05	.24	-.18	—							
10. Immediate Post-Test Similar Score (/12)	52	7.14	2.38	.07	-.05	.00	.05	.00	.08	.32*	-.25	.82***	—						
11. Immediate Post-Test Transfer Score (/8)	52	4.67	2.30	.02	-.01	-.02	.01	.00	.05	.19	-.28*	.80***	.65***	—					
12. Immediate Post-Test Time to Solution (/400)	52	330.79	42.03	.01	.08	.07	.03	.05	.01	-.16	.20	-.65***	-.71***	-.69***	—				
13. Delayed Post-Test Score (/20)	51	13.53	4.22	.02	-.09	.04	-.01	.01	.05	.13	-.18	.50***	.40***	.57***	-.38**	—			
14. Delayed Post-Test Similar Score (/10)	51	7.49	2.06	.04	-.07	.09	.01	.01	.04	.15	-.15	.38**	.35*	.39**	-.22	.82***	—		
15. Delayed Post-Test Transfer Score (/10)	51	6.04	2.63	-.01	-.08	-.01	-.02	-.03	.04	.08	-.15	.46***	.34*	.57***	-.41**	.89***	.47***	—	
16. Delayed Post-Test Time to Solution (/400)	51	308.61	50.81	-.11	-.06	-.16	-.10	-.13	-.19	-.06	.10	-.40**	-.33*	-.50***	.63***	-.61***	-.43**	-.60***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

A similar partial correlational analysis was also conducted on participants' first-stage intrinsic motivation and cognitive load. As participants in the alternating condition also applied the non-alternating tracing strategy in the first-stage learning, only dominant tracing, tracing counts, and finger continued counts were included in the analysis. Similar to previous analysis, the first-stage tracing dosage indicators were included in the analysis. As Table 45 shows, first-stage dominant tracing, tracing counts, and finger continued counts did not correlate with participants' first-stage intrinsic motivation and cognitive load.

A similar partial correlational analysis was also conducted on participants' second-stage intrinsic motivation and cognitive load. The tracing dosage indicators that reflect the sum of the two stages were included in the analysis. As Table 46 shows, all six tracing dosage indicators did not correlate with participants' second-stage intrinsic motivation and cognitive load.

In summary, no significant correlation was found between tracing dosage and alternating group participants' learning performance in the practice phase, the immediate post-test phase, and the delayed post-test phase. No correlation was found on participants' intrinsic motivation and cognitive load.

Table 45

Partial Correlations Between Participants' First-Stage Dominant Tracing, Tracing Counts, Finter Continued Counts, Intrinsic Motivation, and Cognitive Load in the Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	
1. First-Stage Dominant Tracing	52	134.86	76.34	—						
2. First-Stage Tracing Counts	52	7.59	4.36	.91***	—					
3. First-Stage Finger Continued Counts	52	28.87	17.32	.93***	.97***	—				
4. First-Stage Intrinsic Motivation (/8)	52	5.59	1.51	-.10	.02	-.05	—			
5. First-Stage Intrinsic Cognitive Load (/8)	52	1.26	1.31	-.13	-.18	-.21	-.36**	—		
6. First-Stage Extraneous Cognitive Load (/8)	52	1.49	1.20	-.09	-.13	-.16	-.43**	.78***	—	
7. First-Stage Germane Processing (/8)	52	6.24	1.04	-.03	.03	.05	.59***	-.69***	-.66***	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 46

Partial Correlations Between Six Tracing Dosage Factors, Intrinsic Motivation, and Cognitive Load in the Alternating Group in Experiment 3

Variable	*n*	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. Dominant Tracing	52	217.07	116.57	—									
2. Non-dominant Tracing	52	69.33	37.10	.81***	—								
3. Tracing Counts	52	15.99	8.68	.87***	.74***	—							
4. Total Tracing	52	286.40	148.18	.99***	.89***	.87***	—						
5. Finger Alternated Counts	52	16.60	9.70	.84***	.83***	.92***	.87***	—					
6. Finger Continued Counts	52	45.71	25.89	.88***	.73***	.96***	.87***	.90***	—				
7. Intrinsic Motivation (/8)	52	5.35	1.67	-.12	.00	-.07	-.10	-.04	-.15	—			
8. Intrinsic Cognitive Load (/8)	52	0.94	1.09	.01	.08	-.01	.03	.04	-.04	-.04	—		
9. Extraneous Cognitive Load (/8)	52	1.72	1.39	.08	.19	.05	.11	.10	.04	-.33*	.32*	—	
10. Germane Processing (/8)	52	6.05	1.08	-.07	-.12	-.06	-.09	-.06	-.06	.29*	-.31*	-.42**	—

* $p < .05$, ** $p < .01$, *** $p < .001$

9.4 Discussion

Reviewing the results of all three experiments, a general finding was the positive correlation between tracing dosage and participants' learning performance in Experiments 1 and 2, but not Experiment 3. In Experiments 1 and 2, all indicators suggested that more tracing actions resulted in better learning performance. This effect was more pronounced in the practice phase, followed by the immediate post-test, and finally the delayed post-test. In Experiment 1 and Experiment 2, the correlations between dosage indicators and practice phase learning performance were generally higher in magnitude than the correlations between dosage indicators and immediate and delayed post-test learning performance, with some indicators in Experiment 2 in particular not correlating with the post-test learning performance. These findings suggested tracing dosage effects gradually weakened as the time from the learning phase increased.

However, no significant correlation between tracing effect and learning performance was found in Experiment 3. As introduced in the main chapters, Experiment 3 made substantial adjustments to the learning materials and procedure. Both the first-stage practice score and the second-stage practice score evinced ceiling effects, with range restriction of these variates undermining subsequent correlational analyses. It is unclear, however, why dosage effects seen on immediate and delayed post-test scores in Experiments 1 and 2 were not observed on these variates in Experiment 3.

One possible reason is that adjustments to the learning procedure resulted in an extended interval between the acquisition stage and the post-test stage. The findings from Experiments 1 and 2 indicated that the dosage effect diminished rapidly over time. In

Experiment 1 and 2, participants were asked to complete a self-report survey after they answered the fourth practice question. Participants spent approximately five minutes answering the questions in self-report survey, then they started the immediate post-test. A mere five-minute interval resulted in the reduction in the magnitude of dosage effect from practice phase to the immediate post-test phase. Since the learning phase was divided into two stages in Experiment 3, participants were asked to complete two self-report surveys (one after each stage). The time interval between the acquisition stage and immediate-test stage increased due to the procedure's adjustments, associated with a further weakening of the dosage effect in the immediate post-test phase, as well as the delayed post-test.

Though the increased time interval may have reduced the magnitude of the dosage effect, given that the dosage effect was robust in both Experiments 1 and 2, this appears insufficient to account for its complete disappearance in Experiment 3. The reasons for the disappearance of the dosage effect in Experiment 3 require further investigation.

Chapter Ten: General Discussion

10.1 Summary of Key Findings

Based on the tracing effect and drawing on variability's superiority in instructional design as well as CLT's theoretical framework, incorporating variability into which hand performs tracing actions may result in beneficial effects on schema construction and transfer of training (van Merriënboer & Sweller, 2005). Therefore, the current research conducted three experiments in sequence to compare alternating tracing and non-alternating tracing, hypothesising that alternating which hand traces out specified elements of a worked example could further support learning beyond the basic tracing effect. In addition, the series of experiments investigated learners' intrinsic motivation and cognitive load, because the cooperation of tracing and variability may also optimise their respective effects on intrinsic motivation and various sources of cognitive load.

All three experiments recruited college students as participants and used lesson materials on a mental mathematics learning strategy as learning content. Experiment 1 and Experiment 2 employed the same lesson design, with participants studying four worked examples and completing the practice questions that followed each worked example. Contrary to hypotheses, Experiment 1 and 2 did not find that alternating tracing promoted learning compared to non-alternating tracing across all testing phases, including practice questions, immediate post-test questions, and delayed post-test questions. Experiment 1 found non-alternating tracing group performed better than the alternating tracing group in practice questions and immediate post-test questions, with no difference between the groups in delayed post-test questions. The only positive finding regarding alternating tracing in

Experiment 1 was participants in the alternating tracing condition had a greater improvement across immediate post-test and delayed post-test. As for Experiment 2, no difference on learning performance between two groups was found across practice questions, immediate post-test questions, and delayed post-test.

Considering that the action of alternating fingers during learning may cause learners difficulty in focusing on the learning content, Experiment 3 made adjustments to the design of the lesson phase, dividing this phase into two stages, with three worked examples and corresponding practice questions in each stage. In the first stage, participants in both groups applied a non-alternating tracing strategy. This change was implemented in the expectation that it would support initial schema acquisition in both conditions. In the second stage, participants in the non-alternating group continued tracing with the dominant hand, while participants in the alternating tracing group applied the alternating tracing strategy. These modifications were based on the hypothesis that, having partially developed a schema for the mental mathematics strategy in the first phase, students in the alternating condition would further develop and automate the mental mathematics strategy in the second stage to a greater extent than participants in the non-alternating condition, because of the higher (but manageable) intrinsic load and germane processing generated by alternating tracing. In addition to these changes to the lesson phase, immediate and delayed post-test questions were categorised into similar questions and transfer questions, in order to further compare alternating tracing and non-alternating tracing's effects in detail. Thus, the comparison between non-alternating tracing and alternating tracing focused on participants' performance in second-stage practice questions, immediate post-test questions (similar, transfer, and combined test scores), and delayed post-test questions (similar, transfer, and combined test scores). Marginal significant differences between the two groups were found for immediate post-test transfer performance and delayed post-test overall performance, suggesting participants in the alternating group performed better. A marginally significant difference between conditions was also found in the repeated measures analyses across immediate and delayed post-tests, regarding similar question performance, transfer question performance, and combined score performance.

Considerations of intrinsic motivation and cognitive load also yielded significant findings. (Discussion below focuses on the self-report results from Experiment 1, Experiment 2, and the second stage of Experiment 3, because Experiment 3 divided learning into two stages, participants in the alternating condition applied non-alternating tracing in the first stage and applied alternating tracing in the second stage.) For intrinsic motivation, participants in the alternating tracing group reported similar (Experiment 1 and 2) or higher (Experiment 3, stage 2) levels of intrinsic motivation than those in the non-alternating tracing group. More specifically, in Experiment 3, all participants exhibited a decline in intrinsic motivation during the second stage of learning compared to the first, though the decrease was less pronounced for the alternating tracing group. For intrinsic cognitive load, participants in the alternating tracing group reported similar (Experiment 3, stage 2) or higher (Experiment 1) levels of intrinsic cognitive load than those in the non-alternating tracing group. An interaction effect was found in Experiment 1, indicating that participants with lower pre-test scores reported significantly higher intrinsic cognitive load in the alternating condition than those in the non-alternating condition. For extraneous cognitive load, participants in the alternating tracing group reported significantly (Experiment 1 and 2) or marginally (Experiment 3, stage 2) higher levels of extraneous cognitive load than those in the non-alternating tracing group. For germane processing, participants in the alternating tracing group reported similar levels of germane processing to those in the non-alternating tracing group (Experiments 1, 2, and 3).

In addition, supplementary “dosage” analyses of tracing were conducted. Generally, a significant positive correlation between tracing dosage (broadly defined as the number of times a student completed specified tracing actions, reflecting by six indicators) and learning performance was found (Experiments 1 and 2), suggesting that making more tracing actions resulted in better learning performance. Tracing dosage effects were most clear on practice phase scores, less strong on immediate post-test scores, and less strong again on delayed post-test scores. Dosage effects were not apparent in Experiment 3, which might be attributed to ceiling effects in the case of practice test scores. Regarding the lack of a dosage effect on immediate and delayed post-test scores, a possible cause is that the extended interval between acquisition phase and post-test phases may have further diminished the dosage effect.

Though three experiments indicated various results in detail, taken together the following general conclusions can be drawn: compared to non-alternating tracing, a) alternating tracing may cause some difficulty during learning, especially for learners with lower prior ability, reflecting by worse practice question performance (Experiment 1), higher reported intrinsic cognitive load (Experiment 1 and 2) and extraneous cognitive load (Experiment 1, 2, and 3); b) alternating tracing produces better learning retention effects, especially in the delayed post-test and on transfer questions, reflected by greater progress across a 24-hour delay in testing (Experiment 1), better performance in immediate post-test (Experiment 3, transfer scores), and better performance in delayed post-test (Experiment 3, total scores); c) alternating tracing could more effectively sustain students' intrinsic motivation (Experiment 3). Although some of the previous effects are only marginally significant from a statistical perspective, the series of experiments innovatively demonstrates the potential of alternating tracing over non-alternating tracing, further revealing the potential of alternating tracing's application in educational settings.

10.2 Theoretical and Methodological Implications

Based on this series of experiments and research findings, the present study makes the following theoretical contributions.

10.2.1 Exploring Variability from A New Perspective: The Learner's Body as a Learning Medium

First, the present study extends the scope of the variability effect from a CLT perspective. Though various types of variability (for a review, see Raviv et al., 2022) have been studied by researchers from different fields, most of these studies have focused on variation of learning content. Whether variability is operationalised through numerosity, heterogeneity, or scheduling, these factors essentially result in learners acquiring different learning content at specific points in time. Only the variability from situational (contextual) diversity, focusing more on the learning environment, is irrelevant to the learning content.

The present study, however, explores variability from a new perspective. Alternating finger tracing, as a high variability condition compared to non-alternating finger tracing, reflects neither variability in learning materials nor variability in the learning environment, but rather a form of variability of the learning medium (the student's body) – that is, variability in the actions made by students themselves. Learners can either gain the haptic information from the non-alternating finger (single medium, low variability) or alternating fingers (constantly alternating mediums, high variability).

Results in the current studies are consistent with those in the classic variability effect research (e.g., Paas & Van Merriënboer, 1994), satisfying the following characteristics: a) high variability may first hinder learning but then benefit generalisation or transfer (see Experiment 1 and 3); b) high variability may increase intrinsic cognitive load (see Experiment 1 and 2). However, although the current study aligns with the variability effect in terms of the above characteristics, only Experiment 3 actually demonstrated a marginal effect whereby alternating could better promote learning. Experiment 1 and 2 were designed with the intention of eliciting the variability effect and indeed produced changes in cognitive load, yet failed to yield better learning performance. The high variability design had negative consequences in Experiment 1 and 2, which could be regarded as a "reversal variability effect".

As van Merriënboer and Sweller (2005) summarised, instructional designs that incorporate high variability can be achieved via different dimensions, such as the manner in which the task is presented, the saliency of defining characteristics, or the context in which the task is performed. The present study, operationalising variability via alternating tracing actions, replicated the variability effect under specific circumstance (Experiment 3, applying the non-alternating/alternating tracing sequence), extending the potential scope of the variability effect.

10.2.2 Demonstrating Alternating Tracing as A Source of Extraneous Cognitive Load

Second, the present study explores alternating tracing as a new type of tracing, demonstrating it as a potential source of extraneous cognitive load.

Alternating tracing, as a new type of tracing, was introduced in the present study, intending to further support cognitive learning. However, due to the classic tracing effect and variability effect exerting opposing influences on extraneous cognitive load, we could not definitively predict how alternating tracing, as a combination of tracing and variability, would influence extraneous cognitive load. As classic tracing reduces extraneous cognitive load (e.g., Ginns & King, 2021; Tang et al., 2019, Wang et al., 2022; Wang et al., 2025) while variability was predicted to increase it, the present study initially anticipated that these effects would cancel each other out, making the hypothesis that participants in the non-alternating tracing group and alternating tracing group would perceive similar levels of extraneous cognitive load. However, results from all three experiments showed that participants in the alternating tracing group reported higher levels of extraneous cognitive load, preliminarily demonstrating alternating tracing as a source of extraneous cognitive load.

The reason why participants applying alternating tracing reported higher extraneous cognitive load has been discussed in the discussion sections of each experiment. Experiment 1 and Experiment 2 pointed out that there might be a split attention effect (Tarmizi & Sweller, 1988) because participants needed to read the tracing instructions during the learning. However, this effect should exist in both conditions, yet is insufficient to explain why extraneous cognitive load is higher in the alternating tracing group. Although the instructions for alternating tracing are more complex, which may cause a more severe split attention effect, this cannot explain why the higher extraneous cognitive load persisted in the alternating group even after simplifying the instructions to images in Experiment 3. Thus, other potential causes of higher extraneous cognitive load in the alternating tracing group are discussed in the following paragraph. These possibilities may eliminate the possible benefits of heightened attentional resources resulting from finger tracing.

One possible explanation was that having to check carefully to know which hand to use consumed part of working memory resources, resulting in higher extraneous cognitive load. Although both groups of participants followed tracing instructions to use their fingers, participants in the non-alternating tracing group may have developed some automation in this procedure due to consistently using the same hand. In contrast, participants in the alternating tracing group had to pay attention to the tracing instruction at every step, potentially consuming more working memory resources compared to the non-alternating tracing group. Results from Experiment 3 support this explanation. In Experiment 3, participants in both conditions applied a non-alternating tracing strategy in the first-stage learning; those who continuously applied a non-alternating tracing strategy reported lower extraneous cognitive load in the second stage, compared to the first stage. Because participants developed some automation in the process of "checking tracing instruction, using the index finger of the right hand" in the first-stage learning, their extraneous cognitive load may have decreased when they continued to use the right hand in the second-stage learning. In contrast, participants in the alternating tracing group could not continue to rely on this established automation process, resulting in higher extraneous cognitive load in the second stage.

Another possible reason is the difference in ability between the dominant and non-dominant hands. Although finger tracing is a relatively simple hand movement that can be easily performed by the non-dominant hand without specific training, the disparity in proficiency between the use of the dominant and non-dominant hands is an objective reality. Results of the handedness survey supported this disparity. Handedness scores in both groups in all three experiments are higher than 0.8, suggesting that participants had a pronounced preference for using the right hand. The difference in hand preference raises the distinct possibility that the tracing with the non-dominant hand was much less automated than the dominant hand for most participants, resulting in higher extraneous cognitive load in the alternating tracing group.

In summary, the present study introduces alternating tracing, demonstrates its potential effect in increasing extraneous cognitive load, and points out the possible reasons for further consideration.

10.2.3 Constituting Alternating Tracing as A New Compound Cognitive Effect

Third, the present study provides tentative evidence for alternating tracing actions, potentially constituting a new “compound” cognitive load effect alongside the tracing/imagination effect (Wang et al., 2022). As Sweller et al. (2019) defined, a compound effect is an effect that can alter the characteristics of other simple cognitive load effects.

The alternating tracing effect could be seen as a compound effect because it can alter the characteristics of the classic tracing effect. Classic finger tracing was demonstrated to decrease participants’ extraneous cognitive load (e.g., Ginns & King, 2021; Tang et al., 2019; Wang et al., 2025). In marked contrast to these results, results of Experiments 1 and 2 found that alternating tracing increased extraneous cognitive load for novices, compared to standard (non-alternating) tracing.

Following results of Experiments 1 and 2, the redesign of Experiment 3 first aimed to provide all students with foundational schemas through non-alternating tracing actions in the first stage of the lesson, then tested the effect of alternating vs. non-alternating tracing actions in the second stage. Compared to Experiments 1 and 2, differences between conditions in extraneous load were much less pronounced in Experiment 3. Moreover, in Experiment 3, a marginal interaction effect was found for extraneous cognitive load, suggesting the non-alternating group experienced lower extraneous cognitive load than the alternating group in the second stage. However, this seemingly "harmful" alteration resulted was associated with better delayed post-test overall performance and transfer performance, indicating the role of alternating tracing as a compound effect in facilitating learning.

10.2.4 Mitigating the Decline in Intrinsic Motivation via Changing the Tracing Strategy

Fourth, Experiment 3 in the present study demonstrates that applying a non-alternating/alternating tracing sequence could more effectively mitigate the decline in intrinsic motivation across a lesson.

Drawing on evolutionary educational psychology, Ginns and King (2021) proposed finger tracing, as biologically primary knowledge (Geary, 2008), can inherently stimulate humans' motivation to engage in such activities. However, empirical findings regarding finger tracing leading to heightened intrinsic motivation are not robust. Some studies demonstrated that tracing triggered higher motivation compared to the non-tracing group (e.g., Ginns & King, 2021; Wang et al., 2022, Experiment 1), while other studies did not find any difference (e.g., Park et al., 2023; Wang et al., 2022, Experiment 2, 2025). One possible explanation for these various results is that finger tracing did trigger intrinsic motivation in all studies, but some motivation rapidly diminishes during the learning process and cannot be noticed and measured at the time of self-reporting. According to Hidi and Renninger's (2006) four-phase model of interest, triggered situational interest may disappear if the learning lacks novelty. Consistent with this argument, the intrinsic motivation for finger tracing triggered by biologically primary knowledge-related activity may also disappear quickly if the tracing activity or the learning contents lack variation.

Experiment 3 in the present study designed a learning sequence that first applied non-alternating tracing and then applied alternating tracing, successfully mitigating the decline in learners' intrinsic motivation. In contrast, Experiments 1 and 2 applied the alternating tracing across the whole learning process, without finding a difference in intrinsic motivation between the non-alternating and alternating conditions. The series of findings suggests that applying a changeable tracing strategy to maintain learners' sense of novelty in learning activities could be the key factor to preventing learners' intrinsic motivation from declining too rapidly. This discovery holds potential for extension to instructional design more broadly, suggesting that a variety of activities could be applied within the same lesson sequence to maintain students' motivation (particularly interest) while learning biologically secondary knowledge that is not intrinsically motivating to learn (cf. Geary, 2008).

10.2.5 Contributions to Haptic Modality Effect

Fifth, the present study contributes to research on the modality effect via the haptic modality.

The modality effect (Mousavi et al., 1995) originally referred to the finding that dual-modality presentation visual and auditory lesson elements can promote learning. As Sweller et al. (2011) concluded, the modality effect operates by reducing the extraneous cognitive load caused by visual split attention and providing extra working memory resources by adding another sensory channel. By comparing finger tracing on the paper (has haptic perception) and tracing above the paper (no haptic perception), F.-T. Hu et al. (2015) found tracing on the paper could better enhance learning, demonstrating the modality effect exists in the haptic sensory mode.

By adding an extra finger, the present study introduces alternating tracing involving two index fingers in learning. As humans' hands can move independently and each has sensory perception, a question regarding the dual-hands modality arises: does the addition of an extra hand enrich the original haptic modality, or does it constitute its own haptic modality?

Research from neuroscience suggests that both hands possess independent sensory capabilities while also being influenced by one another. Neuroscience has defined a specialised term, proprioception, to describe the perception of movement and spatial orientation arising from stimuli within the body (Brumagne et al., 2013). As proprioception originates from mechanoreceptors within the muscles, tendons, and skin (Casadio et al., 2018), each hand has its own proprioceptive input, which makes the tactile and kinaesthetic perception to some extent independent for each hand. However, numerous studies have also found that tactile sensation undergoes interhemispheric transfer (Tamè et al., 2019). Braun et al.(2005) found that stimuli applied to the left hand substantially altered localization responses for stimuli applied to the right side, suggesting that perception in the left and right hands is to some extent mutually influenced.

Returning to the perspective of CLT on modality effect, the question remains unanswered: whether the left and right hands share the same haptic modality, or do the left and right hand have their own sub-modality? The findings on cognitive load measures in the present study potentially reveal that the processing of tactile and kinaesthetic information by the left and right hands is more like a single modality.

In all three experiments in the present study, the alternating tracing group reported higher extraneous cognitive load. If the information perceived by the left hand and right hand constitutes two relatively distinct haptic modalities, then the findings do not align with the basic idea of the modality effect that more modalities presenting related information results in lower extraneous cognitive load. On the contrary, if the left hand and right hand share a single haptic modality, then the increase of extraneous cognitive load could be explained as a kind of split attention effect (Tarmizi & Sweller, 1988) on the haptic modality. Just as visual material appears momentarily on the left side of the field of vision, then momentarily on the right side of the field of vision, the tactile and kinaesthetic perception shifts back and forth between two hands when applying alternating finger tracing. Learners have to spend extra effort to integrate the tactile and kinaesthetic information perceived by two hands, and thus, the extraneous cognitive load increases.

The direct comparison between non-alternating finger tracing and alternating finger tracing excludes other potential sources of interference that could lead to higher extraneous cognitive load (e.g., visual split attention), because the two conditions used the same learning material and any interference from other modalities should have affected both groups equally.

Under these circumstances, the robust higher extraneous cognitive load reported by the alternating tracing group can only be reasonably explained as the split attention effect on the haptic modality. Thus, the present study contributes to the theory of haptic modality effect, revealing that the two hands basically share the same haptic modality.

10.3 Educational Implications

Finger tracing has extensive application in educational practice. Tracing effects have been found across a range of subject areas (e.g., letters, Bara et al., 2004; operation rules, Ginns et al., 2016; geometry rules, Hu et al., 2015; geometrical shapes, Kalenine et al., 2011; human physiology, Macken & Ginns, 2014; scientific knowledge, Tang et al., 2019), learner demographics (e.g., primary-school-aged children, Agostinho et al., 2015; adults, Ginns & King, 2021), and pedagogical methods (e.g., worked examples, Ginns et al., 2016; expository text with diagrams, Macken & Ginns, 2014; illustrated poster, Tang et al., 2019). As an effective, low-cost, and easy-to-implement learning strategy, finger tracing can be used by students, parents, teachers, and educators. The tracing effect has been demonstrated under a classroom setting (e.g., Du & Zhang, 2019), suggesting finger tracing could be applied in both personal learning and classroom teaching.

Building upon the robust tracing effect, the present study introduces a new tracing strategy: alternating tracing. Experiment 1 and 2 tested the effects of employing alternating tracing across the entire learning process, finding that its learning facilitation effect was equivalent to non-alternating tracing (Experiment 2 and delayed post-test in Experiment 1), and in some cases even inferior to non-alternating tracing (practice test and immediate post-test in Experiment 1). Thus, the present study proposed a learning sequence combining non-alternating tracing followed by alternating tracing in Experiment 3, aiming to provide students with foundational schemas through non-alternating tracing in the first stage of the learning. Results suggested that the non-alternating/alternating sequence could better enhance transfer learning outcomes and delayed post-test performance.

In summary, when taken together, results of Experiments 1 and 2 argue against encouraging learners to alternate which hand traces across the entirety of a set of worked examples. The results of Experiment 3 provide tentative evidence for an effective tracing sequence, where learners first undertake the classic non-alternating tracing before progressing to alternating which hand performs tracing actions. This tracing sequence mitigates the potential adverse impacts of alternating tracing being overly complex during the initial stages of learning.

10.4 Limitations of the Present Study and Suggestions for Future Research

10.4.1 Small Effect Size

The present study introduced alternating tracing as a new tracing strategy, investigating effects on learning performance, intrinsic motivation, and cognitive load. Experiment 1 and 2 did not find alternating tracing was more beneficial for learning than non-alternating tracing. Learning performance results for Experiment 3 were marginally significant in statistical terms. One possible explanation for this modest effect is insufficient sample size.

Initial power analysis for the current study suggested that at least 54 participants for each group and 108 participants in total were required for sufficient experimental power. Although the three experiments recruited participants according to this result and included at least 108 participants in each experiment, excluding participants who did not follow the tracing strategy resulted in a lower number of participants being included in the data analyses. The compliant datasets meeting the criteria in Experiment 1 to 3 comprised 95, 100 and 97 participants respectively, failing to meet the minimum requirement of 108 participants.

Data analyses from Experiment 3 further substantiate this argument. In the analysis of similar question performance in Experiment 3, the ANCOVA analysis focused on immediate post-test similar question performance showed no difference between the alternating and non-alternating group, while the ANCOVA analysis focused on delayed post-test similar question performance also showed no difference between alternating and non-alternating group. However, the repeated measures analysis, including data across both the immediate and delayed post-test, found a marginally significant difference between the alternating and non-alternating groups. As the repeated measures analysis included more data, the analysis had higher statistical power, thereby detecting a marginally significant difference between groups in the repeated measure analysis that would otherwise remain undetectable in the single time-point (i.e., immediate test vs. delayed test) ANCOVA analysis.

Therefore, future research could increase the sample size or ensure recruited participants are more compliant with experimental conditions, in order to support more sensitive tests of hypotheses.

In addition to the issue of effect size, during the process of conducting a series of experiments, some limitations of the present study gradually became apparent. These limitations in research design are listed and discussed below, accompanied by suggestions for improvements in future research.

10.4.2 Unbalanced Gender Distribution

The present study recruited participants in two universities in China, where the gender (male/female) ratio is inherently imbalanced. This imbalance stems from the distinctive academic strengths of these two institutions. Both East China Normal University and South China Normal University are normal universities that were historically designed to train teachers, with both universities excelling in education-related disciplines. Despite their efforts in recent years to develop into comprehensive universities, the prevailing perception among Chinese students and parents has resulted in the majority of their students being female. Against this background, the gender ratios in all three experiments of this study were also imbalanced.

Fisher's exact test (Upton, 1992) was used to check the equivalence in gender proportions in all three experiments. Results showed that gender was evenly distributed in Experiments 1 and 2. However, this was not achieved in Experiment 3. In Experiment 3, more male participants were assigned to the non-alternating condition (12 participants) compared to the alternating condition (four participants). Given the overall scarcity of male participants, this difference was not considered to be a substantive concern. As recent research on gender differences in mathematical performance increasingly indicates that such differences are typically very small (Hyde et al., 1990; Lindberg et al., 2010), the unbalanced gender distributions in the present study do not threaten the research findings.

From a broader perspective, gender effects were found in some cognitive load research (e.g., Castro-Alonso et al., 2019; Gupta et al., 2022), suggesting male and female participants may benefit to varying degrees from the same instructional design. Although it remains unclear whether there are differences in the benefits of finger tracing for individuals of different genders, given that finger tracing constitutes a specialised instructional design involving participants' hand movements and spatial ability, gender may be an influential factor. Therefore, future research should aim for a balanced gender distribution in each condition in order to achieve a more objective, universal result.

10.4.3 Unexpected Ceiling and Floor Effects

The present study faced a few ceiling and floor effects for different self-report measures in three experiments, which is unexpected.

The learning materials are designed for adult learners who have never encountered this mental math method, adapting from previous studies (e.g., Ginns et al., 2019; Smyrnis & Ginns, 2016; Wang et al., 2022). The difficulty level of the questions was drawn from previous studies, deliberately designed and progressively adjusted over the course of three experiment trials. As the materials are inherited from previous research, the presence of ceiling and floor effects is unexpected.

In instructional design studies such as those informed by cognitive load theory, it is vital that the challenge level of lesson materials and test questions be aligned with the current capabilities (particularly prior knowledge) of students (Cooper & Sweller, 1987). Although none of the participants in the present study had prior exposure to this mental math method, their inherent mathematical aptitude may have rendered the learning material overly simplistic for them, resulting the presence of floor effects on intrinsic cognitive load and extraneous cognitive load. The limitation of the present study lies in its failure to account for differences in the inherent mathematical abilities of participants from top universities in China. Though the present study applied statistical methods to avoid the bias caused by the ceiling and floor effects, such limitations at the source of the data should be avoided as far as possible in future research.

In future studies, a pilot test is suggested to be applied to sensitively check the difficulty of the learning material. The pilot test should share the background with formal experiments, recruiting pilot participants from the same pools as the formal experiments, in order to provide more sensitive tests of hypotheses.

10.4.4 Lack of Non-Tracing Group as Control Group

The present study did not set a control (non-tracing) condition in the three experiments. The initial design of the present study was based on the consideration that the tracing effect had been extensively demonstrated, such that our aim was not to replicate the tracing effect but rather to further explore alternating tracing. Therefore, all the efforts were devoted to comparing alternating tracing with non-alternating tracing.

However, the absence of the no-tracing condition makes it difficult to draw a definitive conclusion on alternating tracing’s effect. We cannot know the result that no difference between non-alternating and alternating conditions means no effects of alternating tracing, or it means the alternating tracing has similar effects to non-alternating tracing. Regarding learning performance, we may reasonably infer that alternating tracing also has similar learning enhancement effects, given the robust tracing effects in various situations.

However, in terms of intrinsic motivation and cognitive load, the absence of a comparison with a no-tracing condition may make the conclusions less robust. We found that alternating tracing sometimes results in higher intrinsic cognitive load, but this improvement did not result in the increase of germane processing as expected. However, this comparison pertains specifically to non-alternating tracing. It is indeed possible that alternating tracing improved germane processing compared to the no-tracing condition, but this was not tested in the present series of experiments.

Therefore, it is recommended to include a control group in further research to allow the above comparisons to be made.

10.4.5 Conduct Experiments in a Second-Language Setting

The present study was conducted in a second-language setting. The learning materials used in the present study were written in English. However, all three experiments in the present study were conducted in different universities in China, where Chinese is participants' first language. The decision to use English materials was initially made in order to continue employing previously adopted materials from B. Wang et al. (2022). To avoid ambiguities in the translation process, whilst also taking into account that Chinese university students have studied English for at least ten years and are able to read English in paragraphs, English-written materials were adopted in the present study. To ensure participants gained a full understanding of the experimental tasks, all oral introductory sessions were conducted by the researcher in Chinese.

The research findings indicated that participants generally achieved relatively high scores, with evidence of a ceiling effect, suggesting that the impact of using English-language materials on learning appears to be minimal. However, validity analyses results of self-report measurements across three experiments suggested that even participants with strong second language proficiency encountered ambiguities when completing experiments in a second language setting. In Experiment 1, the item “ICL_5” had a low factor loading, and the item “GCL_2” was not significantly related to the main factor. In Experiment 2, the identical self-report scale was employed; the “ICL_5” still had a low factor loading, but the item “GCL_2” became significantly related to the main factor, though its factor loading was low. The only difference between Experiment 1 and 2 was that the researcher responded to participants’ enquiries about statement meaning with a more content-based rather than word-by-word translation. In Experiment 3, “GCL_2” was carefully rephrased in English to avoid the potential ambiguity caused by Chinese culture and Chinese-based English learning. Results showed that “GCL_2” was significantly correlated to the main factor and had a sufficient factor loading to be included as an indicator. The increasingly specific translation prompts provided by the researcher resulted in continuously optimised validity across three experiments, indicating that even participants with a high level of proficiency in the second language remain susceptible to the influence of their non-native language.

Research on mathematics learning in bilingual environments has found that even in countries with multiple official languages, children who grew up in a multilingual environment still showed better understanding and problem-solving performance for problems stated in their native language (e.g., Bernardo & Calleja, 2005; Van Rinsveld et al., 2016). Van Rinsveld et al. (2016) further demonstrated that although this native language advantage diminished with age (with more years lived in a bilingual environment), this language effect does not disappear even among adult populations. This result aligns with earlier findings that even highly proficient bilinguals remain more accurate and faster at performing arithmetic tasks in their dominant language (Frenck-Mestre & Vaid, 1993; Marsh & Maki, 1976; McClain & Huang, 1982).

The initial decision to use English material in the current study aimed to avoid translation ambiguities. However, this approach gave rise to unforeseen issues, such as the challenges posed by non-native language acquisition for learners and the effects of individual variations in English proficiency.

Thus, it is important for further research to utilise learning materials that are appropriate to the language background of the participants. Even though translating learning materials may carry a certain risk of ambiguity, this would still be more appropriate than having participants learn using a second language.

10.4.6 Double Standards on Learning and Tracing Dosage Exists in Tracing Study

Learning research typically requires participants to complete a certain amount of learning. The present study also stipulated specific time allocations for learning each worked example, yet the actual experiment revealed potential shortcomings in this design within tracing research. Although three experiments in the present study stipulated that participants had two minutes to study each worked example and specified the exact number of times they needed to trace in each step, we did not prescribe how many times they should study the example questions within the two-minute period. We tried to control the total learning time of each participant, but we could not control the actual amount of time they devoted to studying and their learning speed.

The video coding of participants' learning process in the current study revealed that participants' actual tracing behaviours exhibited substantial variation, subsequently framed as "dosage". As the overall learning duration is controlled, such variations in individual learning speeds are typically regarded as individual factors in standard learning experiments and do not affect the results of intergroup comparisons. However, in the tracing experiment, given the significant positive correlation between tracing dosage and learning performance, differences in tracing frequency arising from these variations in learning speed may potentially influence the results. The differences in learning performance between different participants may have not only stemmed from tracing strategy (non-alternating vs. alternating) but may have also arisen from tracing dosage.

The present study indicates the possibility in tracing studies, that learners receive the same amount of time on learning but ultimately generate different learning doses. There exists a double standard in tracing experiments: we want participants to undertake the same amount of learning, and we also want participants to undertake the same amount of tracing. This issue may not affect studies comparing finger tracing with other learning strategies, as they contrast scenarios with and without tracing behaviour. However, it does impact research comparing different tracing strategies (though such research is not common currently).

Thus, for further research wishing to investigate different tracing strategies in depth, we suggest that learning material could be designed in a more structured format. For example, researchers could play a learning video of a fixed duration and ask the learner to complete corresponding tracing actions at the pace dictated by the video's instruction.

10.4.7 Alternating Tracing, Element Interactivity, and Desirable Difficulties

Variability research has in part been informed by the notion of desirable difficulties (Bjork & Bjork, 2011), referring to findings that some instructional designs apparently create difficulty that actually leads to more durable and flexible learning. However, mixed results were achieved when attempting to apply desirable difficulties in tracing practice, yielding both successes and failures, thereby providing valuable insights for future research.

R.A. Bjork (n.d.) pointed out the necessity to introduce difficulties for the learner, and listed several approaches to manipulate the difficulty, including varying the conditions of practice, distributing practice on a given task, reducing feedback to the learner, using tests as learning events, and so on. Some of these approaches are consistent with those employed in variability research and CLT research. Chen et al. (2018) considered the notion of desirable difficulties through the lens of element interactivity, providing a theoretical base indicating when difficulties are and are not desirable. Chen et al. argued that desirable difficulties may be effective for learning low element interactivity information but not high element interactivity information: “additional difficulties that increase element interactivity may not be desirable if element interactivity is already so high that it exceeds working memory capacity. In contrast, increasing element interactivity when it is low may be beneficial provided that the increase in element interactivity does not exceed working memory capacity” (p.6). Difficulties that yield enhanced learning are considered desirable; conversely, difficulties that fail to yield gains are deemed undesirable.

In the current study, alternating tracing constituted a more complex behaviour than non-alternating tracing, with experimental conditions intended to introduce such difficulty. However, in terms of learning performance, Experiments 1 and 2 failed in generating desirable difficulty. After adjusting lesson instructions to have the alternating group use non-alternating tracing before alternating tracing, desirable difficulty appears to have been induced in Experiment 3. Compared to Experiments 1 and 2, the additional difficulty being introduced to the alternating group in Experiment 3 was lower. Participants in alternating groups in Experiments 1 and 2 had to face the new learning content (a mentally demanding mental math strategy) and complex tracing actions (alternating tracing) at the same time across Worked Examples 1- 4. Under a cognitive load analysis, element interactivity was too high in the design of this lesson. Participants were confronted not only with unfamiliar learning material, which in itself placed a significant burden on working memory, but also with unfamiliar tracing methods, resulting in cognitive load exceeding working memory limits. In contrast, in Experiment 3, participants had capacity to process information related to the learning content in the first stage of learning, thereby forming a schema, which subsequently reduced element interactivity for similar content in the second stage of the lesson when alternating tracing was introduced.

Given the results across three experiments, the same alternating tracing strategy yielded divergent outcomes in Experiment 1,2 and Experiment 3 regarding the creation of desirable difficulty. In three experiments, the difficulty of the strategy itself remained constant, while the variation lay in the participants' expertise level when confronting this strategy. In Experiment 1 and 2, participants in the alternating group encountered the alternating tracing as a novice, while in Experiment 3, participants had already acquired preliminary knowledge about the mental math strategy when confronted with the alternating tracing in the second stage learning. This finding raises the possibility that the expertise level affects whether the desirable difficulty could be successfully created.

Findings concerning intrinsic cognitive load in the present study also corroborate this possibility. Experiment 1 found that participants in the alternating group with lower pre-test scores reported significantly higher intrinsic cognitive load. Experiment 2 found that participants in the alternating group generally reported significantly higher intrinsic cognitive load. However, Experiment 3 did not find any different on participants second stage intrinsic cognitive load between alternating tracing and non-alternating tracing group. Experiment 3 also found that participants' intrinsic cognitive load in the second stage was generally lower than those in the first stage, suggesting that first stage learning did help participants form certain schema and elevated their expertise level. These findings suggest that participants with higher ability or those who have already formed certain schema can better confront alternating tracing strategy, rendering this instructional intervention a desirable difficulty.

In fact, the specific intrinsic load reported by participants across the three experiments was quite low. In the case of measuring intrinsic cognitive load using a self-report scale comprising nine statements (scored from 0 to 8), the average ICL scores across three experiments were below 2 points. Even in learning tasks of this nature, where the overall difficulty is not particularly high, employing alternating tracing from the outset would likewise present participants with undesirable difficulty. Thus, further research on alternating tracing could first apply non-alternating tracing before proceeding to alternating tracing.

Sweller (2010) argued that increases in element interactivity associated with high variability will be associated with an increase in germane cognitive load. Although no effect of alternating tracing over non-alternating tracing on germane processing was detected in this study, the enhanced element interactivity and high ability learners unaffected by negative impacts suggest the possibility that more experienced learners may benefit more from alternating tracing. Further research should further explore alternating tracing through the lens of desirable difficulties, focusing on whether alternating tracing produces different effects for novice learners and expert learners. A 2 (non-alternating tracing vs. alternating tracing) x 2 (novice learner vs. expert learner) between-subject design could be applied to test this hypothesis.

10.4.8 Disappearance of Tracing Dosage Effect in Experiment 3

The current study found a strong tracing dosage effect in Experiments 1 and 2, but did not find the dosage effect in Experiment 3. Although Experiment 3 involved certain modifications to the procedure compared to Experiments 1 and 2, the disappearance of this pronounced effect was not anticipated. Whilst the increased time interval may partially account for this outcome, other potential explanations require further investigation.

One possible reason is that tracing dosage and learning performance are not always linearly correlated. Just as some studies have found that there may be an inverted U-shaped relationship between students' study time and their academic performance (e.g., Cooper et al., 2006; Tang & Fu, 2008), a similar non-linear relationship may exist between tracing dosage and learning performance. Once tracing reaches a certain threshold, the additional benefit to learning from further tracing begins to diminish. This may explain why we observed a dosage effect in Experiments 1 and 2, which comprised four worked examples, but did not find this effect in Experiment 3, which contained six worked examples (albeit separated into two stages). Though we have not observed any clear signs of an inverted U-shaped relationship between tracing dosage and learning performance in Experiment 3, the possibility of a non-linear relationship remains a direction worthy of consideration for future tracing dosage research.

Another possible reason is that once extraneous cognitive load experienced during the instructions was reduced, the learning benefits from increased tracing dosage may have been less impactful. The most significant change in materials between Experiment 3 and Experiments 1 and 2 was the replacement of textual tracing instructions with pictorial ones, thereby minimizing extraneous cognitive load arising from reading text. In terms of the potential mechanism of tracing dosage effect, it fundamentally depends on the tracing effect, wherein a qualitative change occurs when the cumulative tracing effect reaches a certain threshold. Given that the tracing effect typically occurs as a reduction in extraneous cognitive load, it is reasonable that the tracing dosage effect ceases to exert its influence in Experiment 3, where extraneous cognitive load was already reduced by learning materials design.

Another suggestion for future tracing dosage research is to stipulate the tracing dosage more strictly. As tracing dosage analysis is a supplementary study within this research, the current study did not establish different conditions based on dosage levels. The dosage analysis arose from diverse tracing behaviours exhibited by participants during the learning phase, thereby constraining the current study to employing partial correlation to explore the relationship between dosage and performance. As it was not originally designed for dosage research, the current study did not strictly restrict the number of tracing actions participants should make. Though the current study asked participants to trace a specified pattern five times in every step and gave them two minutes to learn each worked example, the study did not strictly stipulate how many times the worked example as a whole was studied (including tracing within individual steps) within the two-minute timeframe, which resulted in substantial individual variance in participants' total tracing counts. Future tracing dosage research could directly establish experimental conditions at different levels of tracing dosage to compare different conditions. Regarding the specific regulations on tracing dosage, future research could stipulate that the low dosage group trace each specific pattern five times and the high dosage group trace the same pattern ten times. Participants in both conditions would only make tracing actions when they learned the worked example the first time, and continue learning until two minutes run out (while not tracing in the remaining time). Through such a design, the study both stipulates that participants in different conditions have prescribed but distinct tracing dosages and guarantees that participants across groups received identical learning durations.

Furthermore, different types of tracing dosage could be considered in future research. The current study focused on how many times the learner made the specified tracing actions, but tracing dosage could be measured in multiple ways. For example, Galbraith and Ginns (2023) considered tracing dosage based on the size of tracing actions rather than the number of tracing actions, comparing tracing the same pattern in a larger oval with a smaller oval. Furthermore, the amount of time participants spend on tracing, or how many different elements are traced within the same worked example, could also be considered forms of tracing dosage.

10.5 Conclusion

Drawing on Montessori's (1912) tracing practice and CLT's theoretical framework, the present study tested effects of incorporating variability into finger tracing. The series of experiments provides tentative evidence that alternating tracing can enhance learning beyond the classic tracing effect, provided learners first have an opportunity to trace worked examples without alternating. These findings are consistent with previous tracing studies and the characteristics of variability effects. The combination of tracing and variability constitutes a new compound effect (Sweller et al., 2019), extending instructional design of tracing instructions to a sequence of worked examples that sustains intrinsic motivation while also supporting learning.

References

Aben, B., Stapert, S., & Blokland, A. (2012). About the Distinction between Working Memory and Short-Term Memory. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00301

Abrams, R. A., Davoli, C. C., Du, F., Knapp, W. H., & Paull, D. (2008). Altered vision near the hands. *Cognition*, *107*(3), 1035–1047. https://doi.org/10.1016/j.cognition.2007.09.006

Agostinho, S., Tindall-Ford, S., Ginns, P., Howard, S. J., Leahy, W., & Paas, F. (2015). Giving Learning a Helping Hand: Finger Tracing of Temperature Graphs on an iPad. *Educational Psychology Review*, *27*(3), 427–443. https://doi.org/10.1007/s10648-015-9315-5

Algina, J. (1982). Remarks On The Analysis Of Covariance In Repeated Measures Designs. *Multivariate Behavioral Research*, *17*(1), 117–130. https://doi.org/10.1207/s15327906mbr1701_8

Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. *Spatial Cognition & Computation*, *5*(4), 307–331. https://doi.org/10.1207/s15427633scc0504_2

Alibali, M. W., & DiRusso, A. A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development*, *14*(1), 37–56. https://doi.org/10.1016/S0885-2014(99)80017-3

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology*, *25*(4), 468–523. https://doi.org/10.1006/cogp.1993.1012

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, *15*(6), 593–613. https://doi.org/10.1080/016909600750040571

American Psychological Association. (2018a). Cognitive load. In *APA dictionary of psychology*. https://dictionary.apa.org/cognitive-load

American Psychological Association. (2018b). Learning. In *APA dictionary of psychology*. https://dictionary.apa.org/learning

American Psychological Association. (2018c). Memory. In *APA dictionary of psychology*. https://dictionary.apa.org/memory

Anastasi, A. (1988). *Psychological testing* (6th ed.). Macmillan.

Andre, T. (1990). Type of Inserted Question and the Study-Posttest Delay. *The Journal of Experimental Education*, *58*(2), 77–86.

Andre, T., & Thieman, A. (1988). Level of adjunct question, type of feedback, and learning concepts by reading. *Contemporary Educational Psychology*, *13*(3), 296–307. https://doi.org/10.1016/0361-476X(88)90028-8

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes1. In K. W. Spence & J. T. Spence (Eds), *Psychology of Learning and Motivation* (Vol. 2, pp. 89–195). Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American*, *225*(2), 82–90. https://doi.org/10.1038/scientificamerican0871-82

Baddeley, A. (1986). *Working memory* (pp. xi, 289). Clarendon Press/Oxford University Press.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(Volume 63, 2012), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Psychology Press.

Baddeley, A., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A., & Logie, R. H. (1999). Working memory: The multiple-component model. In *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.005

Bara, F., Gentaz, E., & Colé, P. (2007). Haptics in learning to read with children from low socio-economic status families. *British Journal of Developmental Psychology*, *25*(4), 643–663. https://doi.org/10.1348/026151007x186643

Bara, F., Gentaz, E., Colé, P., & Sprenger-Charolles, L. (2004). The visuo-haptic and haptic exploration of letters increases the kindergarten-children’s understanding of the alphabetic principle. *Cognitive Development*, *19*(3), 433–449. https://doi.org/10.1016/j.cogdev.2004.05.003

Barbieri, C. A., Miller-Cotto, D., Clerjuste, S. N., & Chawla, K. (2023). A Meta-analysis of the Worked Examples Effect on Mathematics Performance. *Educational Psychology Review*, *35*(1), 11. https://doi.org/10.1007/s10648-023-09745-1

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577-+. (WOS:000083668300001). https://doi.org/10.1017/S0140525x99532147

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. (WOS:000253283000023). https://doi.org/10.1146/annurev.psych.59.103006.093639

Beckmann, J. F. (2010). Taming a beast of burden—On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction*, *20*(3), 250–264. https://doi.org/10.1016/j.learninstruc.2009.02.024

Bednarik, R. G. (1986). Parietal Finger Markings in Europe and Australia. *Rock Art Research*, *3*(1), 30–61.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bernardo, A. B. I., & Calleja, M. O. (2005). The Effects of Stating Problems in Bilingual Students' First and Second Languages on Solving Mathematical Word Problems. *The Journal of Genetic Psychology*, *166*(1), 117–129. https://doi.org/10.3200/GNTP.166.1.117-129

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bobis, J., Sweller, J., & Cooper, M. (1993). Cognitive load effects in a primary-school geometry task. *Learning and Instruction*, *3*(1), 1–21. https://doi.org/10.1016/S0959-4752(09)80002-9

Boundy, L., Cameron-Faulkner, T., & Theakston, A. (2019). Intention or Attention Before Pointing: Do Infants' Early Holdout Gestures Reflect Evidence of a Declarative Motive? *Infancy*, *24*(2), 228–248. https://doi.org/10.1111/infa.12267

Box, G. E. P. (1953). Non-Normality and Tests on Variances. *Biometrika*, *40*(3/4), 318–335. https://doi.org/10.2307/2333350

Brady, F. (2004). Contextual Interference: A Meta-Analytic Study. *Perceptual and Motor Skills*, *99*(1), 116–126. https://doi.org/10.2466/pms.99.1.116-126

Braun, C., Hess, H., Burkhardt, M., Wühle, A., & Preissl, H. (2005). The right hand knows what the left hand is feeling. *Experimental Brain Research*, *162*(3), 366–373. https://doi.org/10.1007/s00221-004-2187-4

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology-General*, *136*(4), 539–550. (WOS:000250886700001). https://doi.org/10.1037/0096-3445.136.4.539

Brumagne, S., Dolan, P., & Pickar, J. G. (2013). Chapter 19—What is the relation between proprioception and low back pain? In P. W. Hodges, J. Cholewicki, & J. H. van Dieën (Eds), *Spinal Control* (pp. 219–230). Churchill Livingstone. https://doi.org/10.1016/B978-0-7020-4356-7.00019-7

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A Study of Thinking*. John Wiley and

Sons.

Brünken, R., Plass, J. L., & Moreno, R. (2010). Current Issues and Open Questions in Cognitive Load Research. In J. L. Plass, R. Brünken, & R. Moreno (Eds), *Cognitive Load Theory* (pp. 253–272). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.014

Brysbaert, M. (2019). How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, *2*(1), 16–16. https://doi.org/10.5334/joc.72

Cabrera-Nguyen, P. (2010). Author Guidelines for Reporting Scale Development and Validation Results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research*, *1*(2), 99–103. https://doi.org/10.5243/jsswr.2010.8

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Carroll, W. M. (1994). Using Worked Examples as an Instructional Support in the Algebra Classroom. *Journal of Educational Psychology*, *86*(3), 360–367.

Casadio, M., Iandolo, R., Nataletti, S., Marini, F., Morasso, P., Ponassi, V., & Scheidt, R. A. (2018). Chapter 21—Robotic techniques for the assessment of proprioceptive deficits and for proprioceptive training. In R. Colombo & V. Sanguineti (Eds), *Rehabilitation Robotics* (pp. 289–303). Academic Press. https://doi.org/10.1016/B978-0-12-811995-2.00021-7

Castro-Alonso, J., Wong, A. (Anna), Adesope, O., Ayres, P., & Paas, F. (2019). Gender imbalance in instructional dynamic versus static visualizations: A meta-analysis. *Educational Psychology Review*, *31*, 1–27. https://doi.org/10.1007/s10648-019-09469-1

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, *8*(4), 293–332.

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, *62*(2), 233–246.

Chandler, P., & Sweller, J. (1996). Cognitive Load While Learning to Use a Computer Program. *Applied Cognitive Psychology*, *10*(2), 151–170.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Undesirable Difficulty Effects in the Learning of High-Element Interactivity Materials. *Front Psychol*, *9*, 1483. PubMed-not-MEDLINE (30150964). https://doi.org/10.3389/fpsyg.2018.01483

Chen, O., Kalyuga, S., & Sweller, J. (2016). Relations between the worked example and generation effects on immediate and delayed tests. *Learning and Instruction*, *45*, 20–30. https://doi.org/10.1016/j.learninstruc.2016.06.007

Chen, O., Paas, F., & Sweller, J. (2021). Spacing and Interleaving Effects Require Distinct Theoretical Bases: A Systematic Review Testing the Cognitive Load and Discriminative-Contrast Hypotheses. *Educational Psychology Review*, *33*(4), 1499–1522. https://doi.org/10.1007/s10648-021-09613-w

Chen, O., Paas, F., & Sweller, J. (2023). A Cognitive Load Theory Approach to Defining and

Measuring Task Complexity Through Element Interactivity. *Educational Psychology Review*, *35*(2), 63. https://doi.org/10.1007/s10648-023-09782-w

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, *5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The Role of Gesture in Bilingual Education: Does Gesture Enhance Learning? *International Journal of Bilingual Education and Bilingualism*, *7*(4), 303–319. https://doi.org/10.1080/13670050408667815

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, Including the Special Issue: State of the Art Research into Cognitive Load Theory*, *25*(2), 315–324. https://doi.org/10.1016/j.chb.2008.12.020

Clopper, C. G., & Pisoni, D. B. (2004). Effects of Talker Variability on Perceptual Learning of Dialects. *Language and Speech*, *47*(3), 207–238. https://doi.org/10.1177/00238309040470030101

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Taylor & Francis Group. http://ebookcentral.proquest.com/lib/usyd/detail.action?docID=1222653

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th edn). Routledge.

Colom, R., Shih, P. C., Flores-Mendoza, C., & Quiroga, M. A. (2006). The real relationship between short-term memory and working memory. *Memory (Hove, England)*, *14*(7), 804–813. https://doi.org/10.1080/09658210600680020

Comrey, A. L., & Lee, H. B. (1992). Interpretation and Application of Factor Analytic Results. In *A First Course in Factor Analysis* (2nd edn). Psychology Press.

Congdon, E. L., Kwon, M.-K., & Levine, S. C. (2018). Learning to measure through action and gesture: Children's prior knowledge matters. *Cognition*, *180*, 182–190. https://doi.org/10.1016/j.cognition.2018.07.002

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058. (WOS:000252581900027). https://doi.org/10.1016/j.cognition.2007.04.010

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*(4), 347–362. https://doi.org/10.1037/0022-0663.79.4.347

Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001). Learning by imagining. *Journal of Experimental Psychology-Applied*, *7*(1), 68–82. (WOS:000170951100006). https://doi.org/10.1037/1076-898x.7.1.68

Cooper, H., Robinson, J. C., & Patall, E. A. (2006). Does Homework Improve Academic Achievement? A Synthesis of Research, 1987–2003. *Review of Educational Research*, *76*(1), 1–62. https://doi.org/10.3102/00346543076001001

Cosman, J. D., & Vecera, S. P. (2010). Attention Affects Visual Perceptual Processing Near the Hand. *Psychological Science*, *21*(9), 1254–1258. (WOS:000285454700011).

https://doi.org/10.1177/0956797610380697

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191. https://doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (1995). *Attention and memory: An integrated framework* (pp. xv, 321). Oxford University Press.

Cowan, N. (1999). An Embedded-Processes Model of working memory. In *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/s0140525x01003922

Cowan, N. (2005). *Working memory capacity* (pp. ix, 246). Psychology Press. https://doi.org/10.4324/9780203342398

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Csikszentmihalyi, M., & Hunter, J. (2003). Happiness in Everyday Life: The Uses of Experience Sampling. *Journal of Happiness Studies*, *4*(2), 185–199. https://doi.org/10.1023/A:1024409732742

de Groot, A. D. (1965). *Thought and Choice in Chess*. The Hague: Mouton.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*(3), 292–307. https://doi.org/10.1037/1082-989x.2.3.292

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition* (pp. viii, 413). Harvard University Press.

Du, X., & Zhang, Q. (2019). Tracing worked examples: Effects on learning in geometry. *Educational Psychology (Dorchester-on-Thames)*, *39*(2), 169–187. https://doi.org/10.1080/01443410.2018.1536256

Dubrowski, A., Carnahan, H., & Reznick, R. (2010). Research in Surgical Education: A Primer. In *Key Topics in Surgical Research and Methodology* (pp. 99–114). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71915-1_9

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, *7*(2), 149–158. https://doi.org/10.3758/BF03197595

Eielts, C., Pouw, W., Ouwehand, K., van Gog, T., Zwaan, R. A., & Paas, F. (2020). Co-thought gesturing supports more complex problem solving in subjects with lower visual working-memory capacity. *Psychological Research*, *84*(2), 502–513. Medline (30066133). https://doi.org/10.1007/s00426-018-1065-9

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211

Feldon, D. F., Callan, G., Juth, S., & Jeong, S. (2019). Cognitive Load as Motivational Cost. *Educational Psychology Review*, *31*(2), 319–337.

Flevares, L. M., & Perry, M. (2001). How many do you see? The use of nonspoken representations in first-grade mathematics lessons. *Journal of Educational Psychology*, *93*(2), 330–345. https://doi.org/10.1037/0022-0663.93.2.330

Florax, M., & Ploetzner, R. (2010). What contributes to the split-attention effect? The role of text segmentation, picture labelling, and spatial proximity. *Learning and Instruction*, *20*(3), 216–224. https://doi.org/10.1016/j.learninstruc.2009.02.021

Fodor, J. A. (1975). *The language of thought*. Crowell.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., … R-Core. (2024). *car: Companion to Applied Regression* (Version 3.1-3) [Computer software]. https://cran.r-project.org/web/packages/car/index.html

Frenck-Mestre, C., & Vaid, J. (1993). Activation of number facts in bilinguals. *Memory & Cognition*, *21*(6), 809–818. https://doi.org/10.3758/BF03202748

Galbraith, F., & Ginns, P. (2023). Does the size of tracing actions affect learning outcomes? *Educational and Developmental Psychologist*, *40*(2), 232–243. https://doi.org/10.1080/20590776.2022.2161879

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th edn). Longman.

Geary, D. C. (2007). Educating the evolved mind: Conceptual foundations for an evolutionary educational psychology. In J. S. Carlson & J. R. Levin (Eds), *Educating the Evolved Mind: Conceptual Foundations for an Evolutionary Educational Psychology* (pp. 1–99). Information Age Pub.

Geary, D. C. (2008). An Evolutionarily Informed Education Science. *Educational Psychologist*, *43*(4), 179–195. https://doi.org/10.1080/00461520802392133

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114. https://doi.org/10.1016/j.neuropsychologia.2013.11.010

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, *15*(4), 313–331. (WOS:000232180900003). https://doi.org/10.1016/j.learninstruc.2005.07.001

Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction*, *16*(6), 511–525. https://doi.org/10.1016/j.learninstruc.2006.10.001

Ginns, P., Chandler, P., & Sweller, J. (2003). When imagining information is effective. *Contemporary Educational Psychology*, *28*(2), 229–251. https://doi.org/10.1016/S0361-476X(02)00016-4

Ginns, P., Hu, F.-T., & Bobis, J. (2020). Tracing enhances problem-solving transfer, but without effects on intrinsic or extraneous cognitive load. *Applied Cognitive Psychology*, *34*(6), 1522–1529. https://doi.org/10.1002/acp.3732

Ginns, P., Hu, F.-T., Byrne, E., & Bobis, J. (2016). Learning By Tracing Worked Examples. *Applied Cognitive Psychology*, *30*(2), 160–169. https://doi.org/10.1002/acp.3171

Ginns, P., Kim, T., & Zervos, E. (2019). Chewing Gum While Studying: Effects on Alertness and Test Performance. *Applied Cognitive Psychology*, *33*(2), 214–224. https://doi.org/10.1002/acp.3467

Ginns, P., & King, V. (2021). Pointing and tracing enhance computer-based learning. *Educational Technology Research and Development*, *69*(3), 1387–1403. https://doi.org/10.1007/s11423-021-09997-0

Ginns, P., & Kydd, A. (2019). Learning human physiology by pointing and tracing. In T.-F.

Sharon, A. Shirley, & S. John (Eds), *Advances in Cognitive Load Theory: Rethinking Teaching* (pp. 119–129). Routledge. https://doi.org/10.4324/9780429283895

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*(1), 1–19. https://doi.org/10.1017/S0140525X97000010

Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., & Buccino, G. (2008). Processing abstract language modulates motor system activity. *Quarterly Journal of Experimental Psychology*, *61*(6), 905–919. (WOS:000256760100007). https://doi.org/10.1080/17470210701625550

Gliozzi, V., & Plunkett, K. (2019). Grounding Bayesian accounts of numerosity and variability effects in a similarity-based framework: The case of self-organising maps. *Journal of Cognitive Psychology*, *31*(5–6), 605–618. https://doi.org/10.1080/20445911.2019.1637880

Gobet, F., & Simon, H. A. (1996). Templates in Chess Memory: A Mechanism for Recalling Several Boards. *Cognitive Psychology*, *31*(1), 1–40. https://doi.org/10.1006/cogp.1996.0011

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522. (WOS:000172112700014). https://doi.org/10.1111/1467-9280.00395

Gupta, R., Marcus, N., & Ayres, P. (2022). Investigating the impact of gender-differences and spatial ability on learning from instructional animations. *L'Année Psychologique*, *122*(3), 537–561. https://doi.org/10.3917/anpsy1.223.0537

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, *51*(1), 59–67. https://doi.org/10.1016/S0006-3223(01)01330-0

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, *41*(3), 924–936. https://doi.org/10.3758/BRM.41.3.924

Hidi, S., & Renninger, K. A. (2006). The Four-Phase Model of Interest Development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020). *Environmental drivers of systematicity and generalization in a situated agent*. http://arxiv.org/abs/1910.00571

Hu, F.-T., Ginns, P., & Bobis, J. (2014). Does tracing worked examples enhance geometry learning? *Australian Journal of Educational & Developmental Psychology*, *14*, 45–49.

Hu, F.-T., Ginns, P., & Bobis, J. (2015). Getting the point: Tracing worked examples enhances learning. *Learning and Instruction*, *35*, 85–93. https://doi.org/10.1016/j.learninstruc.2014.10.002

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulme, C. (1979). The interaction of visual and motor memory for graphic forms following tracing. *Quarterly Journal of Experimental Psychology*, *31*(2), 249–261.

https://doi.org/10.1080/14640747908400724

Hulme, C. (1981). The effects of manual tracing on memory in normal and retarded readers: Some implications for multi-sensory teaching. *Psychological Research*, *43*(2), 179–191. https://doi.org/10.1007/BF00309828

Hulme, C., Monk, A., & Ives, S. (1987). Some experimental studies of multi-sensory teaching: The effects of manual tracing on children's paired-associate learning. *British Journal of Developmental Psychology*, *5*(4), 299–307. https://doi.org/10.1111/j.2044-835X.1987.tb01066.x

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139–155. https://doi.org/10.1037/0033-2909.107.2.139

Itoh, S., Morris, T., & Spittle, M. (2022). Examining the frequency variable in the imagery dose-response relationship. *Asian Journal of Sport and Exercise Psychology, Applying Imagery in Sport and Exercise*, *2*(2), 122–130. https://doi.org/10.1016/j.ajsep.2022.06.003

Jensen, N. J., & King, E. M. (1970). Effects of different kinds of visual-motor discrimination training on learning to read words. *Journal of Educational Psychology*, *61*(2), 90–96. https://doi.org/10.1037/h0028913

Jeung, H., Chandler, P., & Sweller, J. (1997). The Role of Visual Indicators in Dual Sensory Mode Instruction. *Educational Psychology*, *17*(3), 329–345. https://doi.org/10.1080/0144341970170307

Johnson, P. O., & Fay, L. C. (1950). The Johnson-Neyman Technique, its Theory and

Application. *Psychometrika*, *15*(4), 349–367. https://doi.org/10.1007/BF02288864

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Julius, E. H. (1992). *Rapid Math Tricks & Tips: 30 Days to Number Power*. Wiley.

Kaas, A. L., Stoeckel, M. C., & Goebel, R. (2008). The neural bases of haptic working memory. In M. Grunwald (Ed.), *Human Haptic Perception: Basics and Applications* (pp. 113–129). Birkhäuser. https://doi.org/10.1007/978-3-7643-7612-3_9

Kalenine, S., Pinet, L., & Gentaz, E. (2011). The visual and visuo-haptic exploration of geometrical shapes increases their recognition in preschoolers. *International Journal of Behavioral Development*, *35*(1), 18–26. https://doi.org/10.1177/0165025410367443

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, *13*(4), 351–371.

Kalyuga, S., & Plass, J. L. (2025). *Rethinking Cognitive Load Theory*. Oxford University Press. https://doi.org/10.1093/9780190078539.001.0001

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, *28*(2), 325–349. https://doi.org/10.1017/S0305000901004664

Kim, H.-Y. (2016). Statistical notes for clinical researchers: Sample size calculation 2. Comparison of two independent proportions. *Restorative Dentistry & Endodontics*, *41*(2), 154–156. https://doi.org/10.5395/rde.2016.41.2.154

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*(4), 351–353. https://doi.org/10.1037/h0046737

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and Gesture* (pp. 162–185). Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.011

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, *48*(1), 45–77. https://doi.org/10.1007/s11251-020-09502-9

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling* (4th edn). Guilford Publications.

Korbach, A., Ginns, P., Brünken, R., & Park, B. (2020). Should learners use their hands for learning? Results from an eye-tracking study. *Journal of Computer Assisted Learning*, *36*(1), 102–113. https://doi.org/10.1111/jcal.12396

Kotovsky, K., Hayes, J. R., & Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology*, *17*(2), 248–294. https://doi.org/10.1016/0010-0285(85)90009-X

Krieglstein, F., Beege, M., Rey, G. D., Sanchez-Stockhammer, C., & Schneider, S. (2023). Development and Validation of a Theory-Based Questionnaire to Measure Different Types of Cognitive Load. *Educational Psychology Review*, *35*(1). https://doi.org/10.1007/s10648-023-09738-0

Krishna, A. (2016). A clearer spotlight on spotlight: Understanding, conducting and reporting. *Journal of Consumer Psychology*, *26*(3), 315–324. https://doi.org/10.1016/j.jcps.2016.04.001

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology*, *70*(2), 144. https://doi.org/10.4097/kjae.2017.70.2.144

Kyun, S., Kalyuga, S., & Sweller, J. (2013). The Effect of Worked Examples When Learning to Write Essays in English Literature. *The Journal of Experimental Education*, *81*(3), 385–408.

Leahy, W., & Sweller, J. (2016). Cognitive load theory and the effects of transient information on the modality effect. *Instructional Science*, *44*(1), 107–123. https://doi.org/10.1007/s11251-015-9362-9

LeBarton, E. S., Goldin-Meadow, S., & Raudenbush, S. (2015). Experimentally-induced Increases in Early Gesture Lead to Increases in Spoken Vocabulary. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, *16*(2), 199–220. https://doi.org/10.1080/15248372.2013.858041

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merrinboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. (WOS:000328272100013). https://doi.org/10.3758/s13428-013-0334-1

Leppink, J., & van den Heuvel, A. (2015). The evolution of cognitive load theory and its application to medical education. *Perspectives on Medical Education*, *4*(3), 119–127. https://doi.org/10.1007/s40037-015-0192-x

Lespiau, F., & Tricot, A. (2022). Primary vs. Secondary knowledge contents in reasoning: Motivated and efficient vs. Overburdened. *Acta Psychologica*, *227*, 1–19. https://doi.org/10.1016/j.actpsy.2022.103610

Levene, H. (1960). Robust Tests for Equality of Variances. In H. Hotelling & I. Olkin (Eds), *Contributions to probability and statistics: Essays in honour of Harold Hotelling* (pp. 278–292). Stanford University Press.

Likourezos, V., Kalyuga, S., & Sweller, J. (2019). The Variability Effect: When Instructional Variability Is Advantageous. *Educational Psychology Review*, *31*(2), 479–497. https://doi.org/10.1007/s10648-019-09462-8

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*, *62*(318), 399–402. https://doi.org/10.2307/2283970

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New Trends in Gender and Mathematics Performance: A Meta-Analysis. *Psychological Bulletin*, *136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Lindemann, O., Stenneken, P., van Schie, H. T., & Bekkering, H. (2006). Semantic activation in action planning. *Journal of Experimental Psychology-Human Perception and Performance*, *32*(3), 633–643. (WOS:000238925400009). https://doi.org/10.1037/0096-1523.32.3.633

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & De Vos, C. (2012). A Prelinguistic Gestural Universal of Human Communication. *Cognitive Science*, *36*(4), 698–713. https://doi.org/10.1111/j.1551-6709.2011.01228.x

Liszkowski, U., Carpenter, M., Henning, A., Striano, T., & Tomasello, M. (2004). Twelve-month-olds point to share attention and interest. *Developmental Science*, *7*(3), 297–307. https://doi.org/10.1111/j.1467-7687.2004.00349.x

Liu, Q., & Wang, L. (2021). T-Test and ANOVA for data with ceiling and/or floor effects. *Behavior Research Methods*, *53*(1), 264–277. https://doi.org/10.3758/s13428-020-01407-2

Longcamp, M., Zerbato-Poudou, M.-T., & Velay, J.-L. (2005). The influence of writing practice on letter recognition in preschool children: A comparison between handwriting and typing. *Acta Psychologica*, *119*(1), 67–79. https://doi.org/10.1016/j.actpsy.2004.10.019

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*, 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Macken, L., & Ginns, P. (2014). Pointing and tracing gestures may enhance anatomy and physiology learning. *Medical Teacher*, *36*(7), 596–601. (WOS:000338197700007). https://doi.org/10.3109/0142159x.2014.899684

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Marquet, J.-C., Freiesleben, T. H., J&oslash, K., Thomsen, R., Murray, A. S., Calligaro, M., Macaire, J.-J., Robert, E., Lorblanchet, M., Aubry, T., Gr&eacute, Bayle, G., Br&eacute, J.-G., h&eacute, ret, Camus, H., Chareille, P., Egels, Y., &Eacute, … Jaubert, J. (2023). The earliest unambiguous Neanderthal engravings on cave walls: La Roche-Cotard, Loire Valley, France. *PLoS ONE*, *18*(6), e0286568–e0286568. https://doi.org/10.1371/journal.pone.0286568

Marsh, L. G., & Maki, R. H. (1976). Efficiency of arithmetic operations in bilinguals as a function of language. *Memory & Cognition*, *4*(4), 459–464. https://doi.org/10.3758/BF03213203

Marshall, S. P. (1995). *Schemas in Problem Solving*. Cambridge University Press. https://doi.org/10.1017/CBO9780511527890

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143

Marton, F. (2015). *Necessary conditions of learning*. Routledge. https://doi.org/10.4324/9781315816876

Marton, F., & Booth, S. (1997). *Learning and awareness*. L. Erlbaum Associates. https://doi.org/10.4324/9780203053690

Matsunaga, M. (2010). How to factor-analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3*(1), Article 1. https://doi.org/10.21500/20112084.854

Mayer, C., Wallner, S., Budde-Spengler, N., Braunert, S., Arndt, P. A., & Kiefer, M. (2020). Literacy Training of Kindergarten Children With Pencil, Keyboard or Tablet Stylus: The Influence of the Writing Tool on Reading and Writing Performance at the Letter and Word Level. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.03054

Mayer, R. E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational Psychology Review*, *8*(4), 357–371. https://doi.org/10.1007/BF01463939

Mayer, R. E. (2001). *Multimedia learning*. Cambridge University Press.

Mayer, R. E. (2002). Rote versus Meaningful Learning. *Theory Into Practice*, *41*(4), 226–232.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, *90*(2), 312–320. https://doi.org/10.1037/0022-0663.90.2.312

Mayer, R. E., & Wittrock, M. C. (1996). Problem-Solving Transfer. In D. C. Berliner & R. C. Calfee (Eds), *Handbook of Educational Psychology* (pp. 47–62). Macmillan.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric Properties of the Intrinsic Motivation Inventory in a Competitive Sport Setting—A Confirmatory Factor-Analysis. *Research Quarterly for Exercise and Sport*, *60*(1), 48–58. (WOS:A1989T494900007). https://doi.org/10.1080/02701367.1989.10607413

McClain, L., & Huang, J. Y. S. (1982). Speed of simple arithmetic in bilinguals. *Memory & Cognition*, *10*(6), 591–596. https://doi.org/10.3758/BF03202441

McDonald, R. P. (1999). *Test theory: A unified treatment* (pp. xi, 485). Lawrence Erlbaum Associates Publishers.

McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research*, *4*(4), 293–307. https://doi.org/10.1007/BF01593882

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/H/bo3641188.html

Meier, S. T., & Feeley, T. H. (2022). Ceiling effects indicate a possible threshold structure for working alliance. *Journal of Counseling Psychology*, *69*(2), 235–245. https://doi.org/10.1037/cou0000564

Mielicki, M. K., & Wiley, J. (2022). Exploring the necessary conditions for observing interleaved practice benefits in math learning. *Learning and Instruction*, *80*, 101583. https://doi.org/10.1016/j.learninstruc.2022.101583

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. Henry Holt and Co. https://doi.org/10.1037/10039-000

Miller, W. A. (1937). The Picture Crutch in Reading. *The Elementary English Review*, *14*(7), 263–274.

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, *22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Montessori, M. (1912). *The Montessori method*. William Heinemann.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, *91*(2), 358–368. https://doi.org/10.1037/0022-0663.91.2.358

Morimoto, T. (2020). The Nature of Haptic Working Memory Capacity and Its Relation to Visual Working Memory. *Multisensory Research*, *33*(8), 837–864. https://doi.org/10.1163/22134808-bja10007

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing Cognitive Load by Mixing Auditory and Visual Presentation Modes. *Journal of Educational Psychology*, *87*(2), 319–334.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity Criteria: Development, Measurement, and Validation. *The American Journal of Evaluation*, *24*(3), 315–340. https://doi.org/10.1177/109821400302400303

Moxley, S. E. (1979). Schema: The variability of practice hypothesis. *Journal of Motor Behavior*, *11*(1), 65–70. PubMed-not-MEDLINE (15186973). https://doi.org/10.1080/00222895.1979.10735173

Nicholls, M. E. R., Thomas, N. A., Loetscher, T., & Grimshaw, G. M. (2013). The Flinders Handedness survey (FLANDERS): A brief measure of skilled hand preference. *Cortex*, *49*(10), 2914–2926. https://doi.org/10.1016/j.cortex.2013.02.002

Nievelstein, F., van Gog, T., van Dijck, G., & Boshuizen, H. P. A. (2013). The worked example and expertise reversal effect in less structured tasks: Learning to reason about legal cases. *Contemporary Educational Psychology*, *38*(2), 118–125. https://doi.org/10.1016/j.cedpsych.2012.12.004

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Oksa, A., Kalyuga, S., & Chandler, P. (2010). Expertise reversal effect in using explanatory notes for readers of Shakespearean text. *Instructional Science*, *38*(3), 217–236.

Olejnik, S. F., & Algina, J. (1984). Parametric ANCOVA and the rank transform ANCOVA when the data are conditionally non-normal and heteroscedastic. *Journal of Educational Statistics*, *9*(2), 129–149. https://doi.org/10.2307/1164717

Owens, P., & Sweller, J. (2008). Cognitive load theory and music instruction. *Educational Psychology*, *28*(1), 29–45. https://doi.org/10.1080/01443410701369146

ÖZTUNA, D., ELHAN, A., & TÜCCAR, E. (2006). Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions. *Turkish Journal of Medical Sciences*, *36*(3), 171–176. https://doi.org/-

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. (WOS:000181088900001). https://doi.org/10.1207/S15326985ep3801_1

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science*, *32*(1–2), 1–8. (WOS:000220172200001). https://doi.org/10.1023/B:TRUC.0000021806.17516.d0

Paas, F., & Sweller, J. (2012). An Evolutionary Upgrade of Cognitive Load Theory: Using the Human Motor System and Collaboration to Support the Learning of Complex Cognitive Tasks. *Educational Psychology Review*, *24*(1), 27–45. https://doi.org/10.1007/s10648-011-9179-2

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122–133. https://doi.org/10.1037/0022-0663.86.1.122

Park, B., Korbach, A., Ginns, P., & Brünken, R. (2019). Embodied cognition? Effects of pointing and tracing gestures on learning performance, eye movement and cognitive load. In S. A. Sharon Tindall-Ford John Sweller (Ed.), *Advances in Cognitive Load Theory: Rethinking Teaching* (pp. 142–154). Routledge. https://doi.org/10.4324/9780429283895-12

Park, B., Korbach, A., Ginns, P., & Brünken, R. (2023). How Learners Use Their Hands for Learning: An Eye-Tracking Study. *Educational Psychology Review*, *35*(4), 116. https://doi.org/10.1007/s10648-023-09833-2

Paulo, F. C., & Robert, L. G. (2014). Effects of Interleaved and Blocked Study on Delayed Test of Category Learning Generalization. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00936

Payne, M., Stringer, D., Carter, B., Hardy, A., & Emsley, R. (2025). Reviewing methodological approaches to dose-response modelling in complex interventions: Insights and perspectives. *BMC Medical Research Methodology*, *25*(1), 135. https://doi.org/10.1186/s12874-025-02585-3

(PDF) Memory and Meta-memory Considerations in the Training of Human Beings. (n.d.). In *ResearchGate*. https://doi.org/10.7551/mitpress/4561.003.0011

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Perry, M., Berch, D., & Singleton, J. (1995). Constructing shared understanding: The role of nonverbal input in learning contexts. *The Journal of Contemporary Legal Issues*, *6*(1), 213.

Peterson, L. R., & Peterson, M. J. (1959). Short-Term Retention of Individual Verbal Items. *Journal of Experimental Psychology*, *58*(3), 193–198. (WOS:A1959WF60300001). https://doi.org/10.1037/h0049234

Phillips, A. (1997). Design and Analysis of Dose-Response Studies: Reality versus Regulatory Requirements. *Drug Information Journal*, *31*(3), 737–744. https://doi.org/10.1177/009286159703100314

Piaget, J. (1952). *The origins of intelligence in children* (p. 419). W. W. Norton & Company. https://doi.org/10.1037/11494-000

Pillay, H. K. (1994). Cognitive load and mental rotation: Structuring orthographic projection for learning and problem solving. *Instructional Science*, *22*(2), 91–113. https://doi.org/10.1007/BF00892159

Ping, R., & Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science*, *34*(4), 602–619. PubMed-not-MEDLINE (21564226). https://doi.org/10.1111/j.1551-6709.2010.01102.x

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*(1), 144–161.

(WOS:A1996UD48700012). https://doi.org/10.1037/0022-0663.88.1.144

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. Medline (35577719). https://doi.org/10.1016/j.tics.2022.03.007

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Reed, C. L., Grubb, J. D., & Steele, C. (2006). Hands up: Attentional prioritization of space near the hand. *Journal of Experimental Psychology-Human Perception and Performance*, *32*(1), 166–177. (WOS:000235435000014). https://doi.org/10.1037/0096-1523.32.1.166

Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Version 2.4.12) [Computer software]. https://cran.r-project.org/web/packages/psych/index.html

Rickards, J. P. (1976). Interaction of Position and Conceptual Level of Adjunct Questions on Immediate and Delayed Retention of Text. *Journal of Educational Psychology*.

Rosenthal, R., & Fode, K. L. (1963). The Effect of Experimenter Bias on the Performance of the Albino Rat. *Behavioral Science*, *8*(3), 183–189.

Rosseel, Y., Jorgensen, T. D., Wilde, L. D., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Rockwood, N., Scharf, F., Du, H., Jamil, H., & Classe, F. (2024). *lavaan: Latent Variable Analysis* (Version 0.6-19) [Computer software]. https://cran.r-project.org/web/packages/lavaan/index.html

Rourke, A., & Sweller, J. (2009). The worked-example effect using ill-defined problems: Learning to recognise designers' styles. *Learning and Instruction*, *19*(2), 185–199. https://doi.org/10.1016/j.learninstruc.2008.03.006

Rueschemeyer, S. A., Lindemann, O., van Elk, M., & Bekkering, H. (2009). Embodied cognition: The interplay between automatic resonance and selection-for-action mechanisms. *European Journal of Social Psychology*, *39*(7), 1180–1187. (WOS:000272572300010). https://doi.org/10.1002/ejsp.662

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Schnotz, W., Fries, S., & Horz, H. (2009). Motivational aspects of cognitive load theory. In M. Wosnitza, S. A. Karabenick, A. Efklides, & P. Nenniger (Eds), *Contemporary motivation research: From global to local perspectives* (pp. 69–96). Hogrefe & Huber. https://plus.cobiss.net/cobiss/si/sl/bib/40623202

Schweppe, J., & Rummer, R. (2014). Attention, Working Memory, and Long-Term Memory in Multimedia Learning: An Integrated Perspective Based on Process Models of Working Memory. *Educational Psychology Review*, *26*(2), 285–306.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sharpe, K., & Van Gelder, L. V. (2006). Evidence for cave marking by Palaeolithic children. *Antiquity*, *80*(310), 937–948.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Simion, F., & Giorgio, E. D. (2015). Face perception and processing in early infancy: Inborn predispositions and developmental changes. *Frontiers in Psychology*, *6*, 969. https://doi.org/10.3389/fpsyg.2015.00969

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*(2), 833–870. Medline (17586482). https://doi.org/10.1016/j.cognition.2007.05.002

Smyrnis, E., & Ginns, P. (2016). Does a Drama-Inspired 'Mirroring' Exercise Enhance Mathematical Learning? *The Educational and Developmental Psychologist*, *33*(2), 178–186. https://doi.org/10.1017/edp.2016.17

Sowinski, C., Dunbar, K., & LeFevre, J. (2014). *Calculation Fluency Test*. Retrieved from https://carleton.ca/cacr/math-lab/measures/calculation-fluency-test/

Spillers, G., Brewer, G., & Unsworth, N. (2012). Working Memory and Information Processing. In *Encyclopedia of the Sciences of Learning* (pp. 3474–3476). Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-1428-6_787

Sundararajan, N., & Adesope, O. (2020). Keep it Coherent: A Meta-Analysis of the Seductive Details Effect. *Educational Psychology Review*, *32*(3), 707–734.

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (1989). Cognitive technology: Some procedures for facilitating learning and problem solving in mathematics and science. *Journal of Educational Psychology*, *81*(4), 457–466. https://doi.org/10.1037/0022-0663.81.4.457

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J. (1999). *Instructional Design*. ACER Press.

Sweller, J. (2003). Evolution of human cognitive architecture. In *The psychology of learning and motivation: Advances in research and theory, Vol. 43* (pp. 215–266). Elsevier Science.

Sweller, J. (2004). Instructional Design Consequences of an Analogy between Evolution by Natural Selection and Human Cognitive Architecture. *Instructional Science*, *32*(1), 9–31. https://doi.org/10.1023/B:TRUC.0000021808.72598.4d

Sweller, J. (2008). Instructional Implications of David C. Geary's Evolutionary Educational Psychology. *Educational Psychologist*, *43*(4), 214–216. https://doi.org/10.1080/00461520802392208

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J. (2012). Human cognitive architecture: Why some instructional procedures work and others do not. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds), *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues.* (pp. 295–325). American Psychological Association.

https://doi.org/10.1037/13273-011

Sweller, J. (2022). The Role of Evolutionary Psychology in Our Understanding of Human Cognition: Consequences for Cognitive Load Theory and Instructional Procedures. *Educational Psychology Review*, *34*(4), 2229–2241. https://doi.org/10.1007/s10648-021-09647-0

Sweller, J. (2023). The Development of Cognitive Load Theory: Replication Crises and Incorporation of Other Theories Can Lead to Theory Expansion. *Educational Psychology Review*, *35*(4), 95. https://doi.org/10.1007/s10648-023-09817-2

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. Springer. Library Hub Discover. http://doi.org/10.1007/978-1-4419-8126-4

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*(3), 185–233. https://doi.org/10.1207/s1532690xci1203_1

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General*, *119*(2), 176–192. https://doi.org/10.1037/0096-3445.119.2.176

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89. https://doi.org/10.1207/s1532690xci0201_3

Sweller, J., & Sweller, S. (2006). Natural Information Processing Systems. *Evolutionary Psychology*, *4*(1). https://doi.org/10.1177/147470490600400135

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, *31*(2), 261–

292. https://doi.org/10.1007/s10648-019-09465-5

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review*, *10*(3), 251–296.

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental Designs Using ANOVA*. Brooks/Cole.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using Multivariate Statistics: Pearson New International Edition*. Pearson.

Tamè, L., Azañón, E., & Longo, M. R. (2019). A Conceptual Model of Tactile Processing across Body Features of Size, Shape, Side, and Spatial Location. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00291

Tang, L., & Fu, L. (2008). An empirical study of relationship between schoolwork burden and academic achievements. *Frontiers of Education in China*, *3*(4), 504–515. https://doi.org/10.1007/s11516-008-0033-3

Tang, M., Ginns, P., & Jacobson, M. J. (2019). Tracing Enhances Recall and Transfer of Knowledge of the Water Cycle. *Educational Psychology Review*, *31*(2), 439–455. https://doi.org/10.1007/s10648-019-09466-4

Tarmizi, R. A., & Sweller, J. (1988). Guidance during Mathematical Problem Solving. *Journal of Educational Psychology*, *80*(4), 424–436.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848. https://doi.org/10.1002/acp.1598

The Intellectus Statistics Team. (2021). *Intellectus Statistics* (Version 2.6) [Computer software]. https://www.intellectusstatistics.com/

The jamovi project. (2024). *Jamovi* (Version 2.6) [Computer software]. https://www.jamovi.org

Thomas, T., & Sunny, M. M. (2019). Situational Determinants of Hand-Proximity Effects. *Collabra-Psychology*, *5*(1). (WOS:000493294200001). https://doi.org/10.1525/collabra.198

Tindall-Ford, S., Chandler, P., & Sweller, J. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied*, *3*(4), 257–287. https://doi.org/10.1037/1076-898X.3.4.257

Trafton, J. G., & Reiser, B. J. (1993). The Contributinos of Studying Examples and Solving Problems to Skill Acquisition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *15*(0). https://escholarship.org/uc/item/71d733c9

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Upton, G. J. G. (1992). Fisher's Exact Test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *155*(3), 395–402. https://doi.org/10.2307/2982890

Van Gelder, L. V. (2015). The Role of Children in the Creation of Finger Flutings in Koonalda Cave, South Australia. *Childhood in the Past*, *8*(2), 149–160. https://doi.org/10.1179/1758571615Z.00000000036

van Merriënboer, J. J. G., & Sluijsmans, D. M. A. (2008). Toward a Synthesis of Cognitive Load Theory, Four-Component Instructional Design, and Self-Directed Learning. *Educational Psychology Review*, *21*(1), 55–66. https://doi.org/10.1007/s10648-008-9092-5 van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177. (WOS:000229220900003). https://doi.org/10.1007/s10648-005-3951-0

Van Rinsveld, A., Schiltz, C., Brunner, M., Landerl, K., & Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter? *Learning and Instruction*, *42*, 72–82. https://doi.org/10.1016/j.learninstruc.2016.01.003

Vukatana, E., Graham, S. A., Curtin, S., & Zepeda, M. S. (2015). One is Not Enough: Multiple Exemplars Facilitate Infants' Generalizations of Novel Properties. *Infancy*, *20*(5), 548–575. https://doi.org/10.1111/infa.12092

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition*, *40*(5), 703–716. https://doi.org/10.3758/s13421-011-0180-2

Walshe, K., Nowell, A., & Floyd, B. (2024). Finger Fluting in Prehistoric Caves: A Critical Analysis of the Evidence for Children, Sexing and Tracing of Individuals. *Journal of Archaeological Method and Theory*, *31*(3), 1522–1542. https://doi.org/10.1007/s10816-024-09646-9

Walton, G. E., Bower, N. J. A., & Bower, T. G. R. (1992). Recognition of familiar faces by newborns. *Infant Behavior and Development*, *15*(2), 265–269. https://doi.org/10.1016/0163-6383(92)80027-R

Wang, B., Ginns, P., & Mockler, N. (2022). Sequencing Tracing with Imagination. *Educational Psychology Review*, *34*(1), 421–449. https://doi.org/10.1007/s10648-021-

09625-6

Wang, Y., Han, K., & Ginns, P. (2025). Tracing and Pointing Support Multimedia Learning: A Cross-Cultural Replication. *Educational Psychology Review*, *37*(2), 32. https://doi.org/10.1007/s10648-025-10005-7

Ward, M., & Sweller, J. (1990). Structuring Effective Worked Examples. *Cognition and Instruction*, *7*(1), 1–39.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Yeo, L.-M. (2024). Mathematics Worked Examples with Tracing Gesture Makes Learning Last. *International Journal of Advanced Research in Education and Society*, *6*(3), Article 3.

Yeo, L.-M., & Tzeng, Y.-T. (2020). Cognitive Effect of Tracing Gesture in the Learning from Mathematics Worked Examples. *International Journal of Science and Mathematics Education*, *18*(4), 733–751. https://doi.org/10.1007/s10763-019-09987-y

Youssef-Shalala, A., Ayres, P., Schubert, C., & Sweller, J. (2014). Using a general problem-solving strategy to promote transfer. *Journal of Experimental Psychology: Applied*, *20*(3), 215–231. https://doi.org/10.1037/xap0000021

Zhang, S., de Koning, B. B., & Paas, F. (2023). Effects of finger and mouse pointing on learning from online split-attention examples. *The British Journal of Educational Psychology*, *93 Suppl 2*, 287–304. https://doi.org/10.1111/bjep.12556

Zhu, X., & Simon, H. A. (1987). Learning Mathematics from Examples and by Doing. *Cognition and Instruction*, *4*(3), 137–166.

Zuo, G., de Koning, B. B., & Paas, F. (2025). Effects of finger tracing on learning from spatially separated and integrated multimedia materials. *Educational Psychology*, *45*(7), 749–765. https://doi.org/10.1080/01443410.2025.2516549

Zuo, G., & Lin, L. (2025). Can whole-body tracing and hand tracing make any difference? Experimental evidence of learning outcomes, cognitive load, and intrinsic motivation on university students. *Instructional Science*, *53*(1), 1–25. https://doi.org/10.1007/s11251-024-09664-w

Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology-General*, *135*(1), 1–11. (WOS:000235388100001). https://doi.org/10.1037/0096-3445.135.1.1

Appendix

A. Full Set of Experimental Materials (Experiment 1)

Demographic Page (1 page)

Participant No. ___________
Date:

Welcome and General Instruction Phase

This experiment is conducted by Xufei Zhang, Ph.D. student from The University of Sydney.

The experiment is approved by the Human Research Ethics Committee of University of Sydney.

The study is about learning to mental math. The following parts of the study will be presented on the paper. Please read the instruction carefully and complete the tests as required.

Demographic Question Phase

Please answer the following questions

1) What is your age? ____________ years

2) What is your gender? (Please circle one answer.)

A. Female B. Male C. Other (please specify) _____________ D. Prefer not to say

Please choose your preferred hand in the following situations

Please tick one box for each question, indicating whether you prefer to use the left-hand, either-hand, or the right-hand for that task. Only tick the 'either' box if one hand is truly no better than the other. Please answer all questions, and even if you have had little experience in a particular task, try imagining doing that.

		Left	**Either**	**Right**
1	With which hand do you write?			
2	In which hand do you prefer to use a spoon when eating?			
3	In which hand do you prefer to hold a toothbrush when cleaning your teeth?			
4	In which hand do you hold a match when you strike it?			
5	In which hand do you prefer to hold the rubber when erasing a pencil mark?			
6	In which hand do you hold the needle when you are sewing?			
7	When buttering bread, which hand holds the knife?			
8	In which hand do you hold a hammer?			
9	In which hand do you hold the peeler when peeling an apple?			
10	Which hand do you use to draw?			

Pre-test Page (3 pages)

Addition

<u>Part 1 (1 minute)</u>

28 +13	51 +10	42 +53	71 +11	95 +52	74 +38	14 +19	99 +63	57 +83	17 +39
19 +27	40 +44	67 +38	98 +59	42 +13	17 +19	90 +82	45 +91	55 +58	83 +42
98 +31	34 +22	20 +54	63 +92	40 +59	26 +89	18 +39	27 +36	44 +80	88 +77
47 +17	23 +48	41 +53	47 +85	59 +16	23 +18	87 +58	31 +53	38 +49	34 +78
41 +38	86 +93	58 +34	25 +77	86 +55	29 +22	74 +31	34 +19	15 +26	83 +19
37 +98	13 +87	38 +67	51 +65	78 +45	89 +32	34 +65	56 +45	23 +43	47 +39

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

Subtraction

Part 2 (1 minute)

89 -60	52 -48	60 -39	51 -28	85 -23	18 -11	49 -37	83 -57	42 -23	68 -47
52 -19	91 -23	60 -31	42 -31	94 -45	98 -64	50 -33	53 -19	61 -45	41 -27
39 -23	61 -37	29 -19	31 -14	54 -12	92 -65	60 -43	43 -27	70 -31	94 -24
48 -19	42 -31	95 -65	81 -62	40 -31	51 -27	42 -18	97 -18	93 -45	74 -23
65 -39	36 -22	80 -46	51 -27	82 -31	91 -64	68 -59	75 -34	64 -24	42 -37
99 -45	61 -27	91 -60	82 -47	25 -19	73 -45	58 -32	57 -17	59 -42	31 -27

DO NOT GO BACK AND

DO NOT TURN THE PAGE UNTIL ASKED TO DO SO.

Multiplication

Part 3 (1 minute)

73 x 8	41 x 5	69 x 3	29 x 9	16 x 8	63 x 8	60 x 4	52 x 4	85 x 6	36 x 7
52 x 9	98 x 3	41 x 8	19 x 6	15 x 4	49 x 2	71 x 9	30 x 8	48 x 7	81 x 5
45 x 9	32 x 6	79 x 2	37 x 8	19 x 9	52 x 6	17 x 5	47 x 2	39 x 3	78 x 7
14 x 3	38 x 8	50 x 2	80 x 9	61 x 7	52 x 7	97 x 4	72 x 7	16 x 5	49 x 6
52 x 7	71 x 2	96 x 8	47 x 6	83 x 3	16 x 8	44 x 3	50 x 8	62 x 6	39 x 5
13 x 8	68 x 4	75 x 5	67 x 8	45 x 9	94 x 7	52 x 6	83 x 7	61 x 6	54 x 9

DO NOT GO BACK TO ANY OTHER PAGE

Learning Page (Non-Alternating Condition, 10 pages)

Acquisition phase

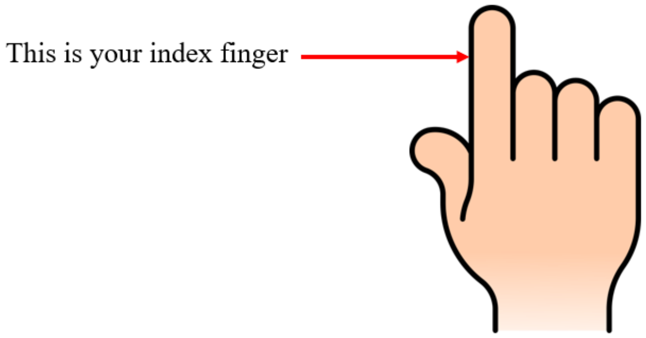
You will now be shown 4 worked examples to study. Please **use your index finger of your writing hand** to help you learn. For each worked example, you will have 2 minutes to:

(1) Look at the worked example;

(2) Read the solution steps in the worked example carefully;

(3)As the instructions in the brackets tell you, use your index finger to trace the operation oval in the worked example on the page.

Please make sure you concentrate on this task because you will be given a very similar problem to solve immediately after each worked example.



Tracing Practice

Please trace out the circle **<u>five times</u>** with your index finger of your **<u>writing hand</u>**.

Please trace out the circle **<u>five times</u>** with your index finger of your **<u>non-writing hand</u>**.

Worked Example 1: 32 x 23 = ?

```
  3 2
x 2 3
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 1. Multiply the ones digits: **2 x 3 = 6** (ones-digit answer).

```
  3  2
x 2  3
     6
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 2. Cross-multiply and add: **(3x3)** + **(2x2)** = 13 (tens-digit answer).
use the 3 as the tens-digit answer then carry the 1.

```
   3  2
x  2  3
 1 3  6
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 3. Multiply and add: (3x2) + 1carried= 7 (hundreds-digit answer).

```
   3  2
x  2  3
7  3  6
```

Answer: 736

Practice Question 1

52 x 21 = ?

Answer: _________

Worked Example 2: 25 x 41 = ?

```
 2 5
x4 1
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 1. Multiply the ones digits: **5 x 1 = 5** (ones-digit answer).

```
    2  5
x   4  1
-------
       5
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 2. Cross-multiply and add: **(2x1)** + **(5x4)** = 22 (tens-digit answer). use the 2 as the tens-digit answer then carry the 2.

```
    2  5
x   4  1
-------
  2 2  5
```

Trace out the oval ***five times*** with your index finger of your ***writing*** hand.

Step 3. Multiply and add: (2x4) + 2carried= 10 (hundreds-digit answer).

```
      2  5
x     4  1
---------
1  0  2  5
```

Answer: 1025

Practice Question 2

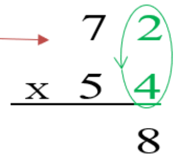
42 x 42 = ?

Answer: _________

Worked Example 3: 72 x 54 = ?

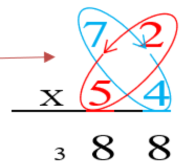
```
  7 2
 x5 4
```

Step 1. Multiply the ones digits: **2 x 4 = 8** (ones-digit answer).



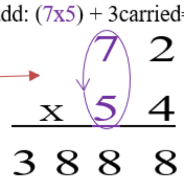
Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 2. Cross-multiply and add: **(7x4)** + **(2x5)** = 38 (tens-digit answer).
use the 8 as the tens-digit answer then carry the 3.



Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 3. Multiply and add: (7x5) + 3carried= 38 (hundreds-digit answer).



Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Answer: 3888

Practice Question 3

53 x 33 = ?

Answer: _________

Worked Example 4: 67 x 23 = ?

```
  6 7
x 2 3
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 1. Multiply the ones digits: **7 x 3 = 21** (ones-digit answer).
use the 1 as the ones-digit answer then carry the 2.

```
    6  7
x   2  3
-------
     2 1
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 2. Cross-multiply and add: **(6x3)** + **(7x2)** = **32** (tens-digit answer).
use the 2+2carried as the tens-digit answer then carry the 3.

```
    6  7
x   2  3
-------
  3 4  1
```

Trace out the oval ***five times*** with your index finger of your ***writing*** hand.

Step 3. Multiply and add: (6x2) + 3carried= 15 (hundreds-digit answer).

```
      6  7
  x   2  3
  --------
1 5   4  1
```

Answer: 1541

Practice Question 4

86 x 43 = ?

Answer: _________

Learning Page (Alternating Condition, 10 pages)

Acquisition phase

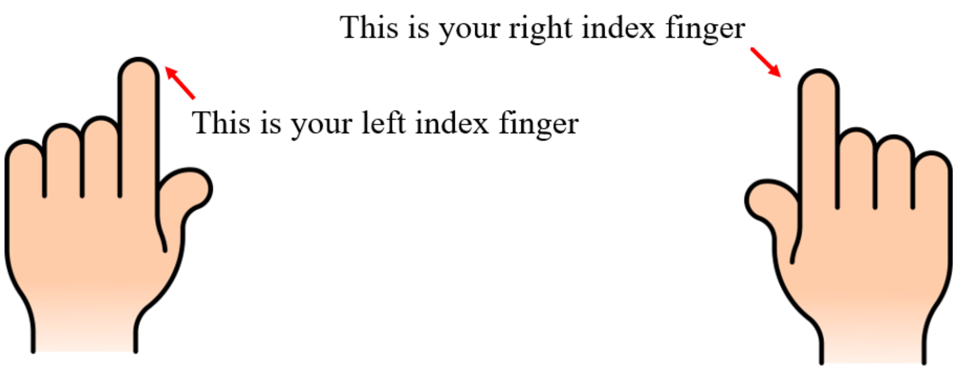
You will now be shown 4 worked examples to study. please use **your finger of your writing or non-writing hand** to help you learn. For each worked example, you will have 2 minutes to:

(1) Look at the worked example;

(2) Read the solution steps in the worked example carefully;

(3) As the instructions in the brackets tell you, use your index finger of your hands to trace the operation oval in the worked example on the page.

Please make sure you concentrate on this task because you will be given a very similar problem to solve immediately after each worked example.



Tracing Practice

Please trace out the circle **<u>five times</u>** with your index finger of your **<u>writing hand</u>**.

Please trace out the circle **<u>five times</u>** with your index finger of your **<u>non-writing hand</u>**.

Worked Example 1: 32 x 23 = ?

```
  3 2
x 2 3
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 1. Multiply the ones digits: **2 x 3 = 6** (ones-digit answer).

```
  3 2
x 2 3
    6
```

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.

Step 2. Cross-multiply and add: **(3x3)** + **(2x2)** = 13 (tens-digit answer).
use the 3 as the tens-digit answer then carry the 1.

```
   3 2
 x 2 3
 1 3 6
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 3. Multiply and add: (3x2) + 1carried= 7 (hundreds-digit answer).

```
  3 2
x 2 3
7 3 6
```

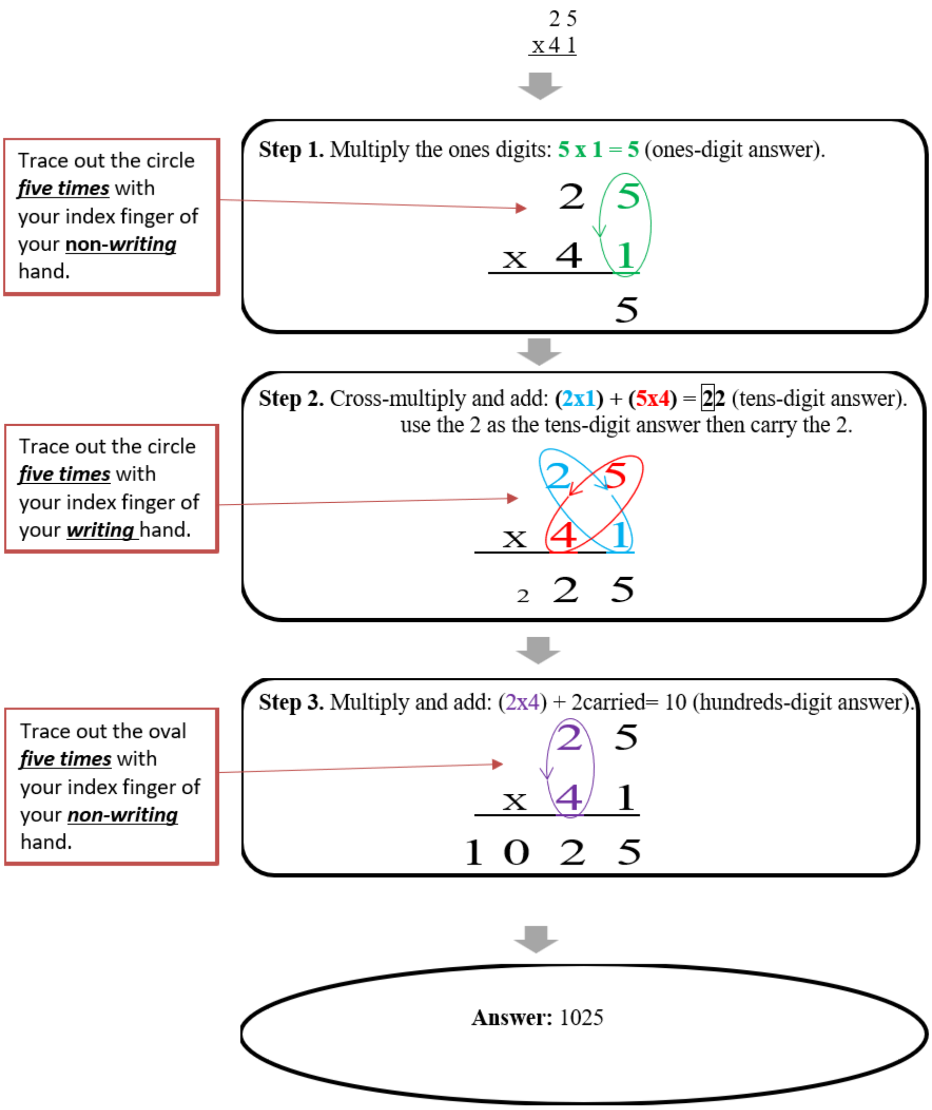
Answer: 736

Practice Question 1

52 x 21 = ?

Answer: _________

Worked Example 2: 25 x 41 = ?



Practice Question 2

42 x 42 = ?

Answer: _________

Worked Example 3: 72 x 54 = ?

```
  7 2
x 5 4
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 1. Multiply the ones digits: **2 x 4 = 8** (ones-digit answer).

```
  7 2
x 5 4
    8
```

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.

Step 2. Cross-multiply and add: **(7x4)** + **(2x5)** = **38** (tens-digit answer). use the 8 as the tens-digit answer then carry the 3.

```
  7 2
x 5 4
 3 8 8
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 3. Multiply and add: (7x5) + 3carried= 38 (hundreds-digit answer).

```
    7 2
  x 5 4
3 8 8 8
```

Answer: 3888

Practice Question 3

53 x 33 = ?

Answer: _________

Worked Example 4: 67 x 23 = ?

```
 6 7
x2 3
```

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.

Step 1. Multiply the ones digits: **7 x 3 = 21** (ones-digit answer).

use the 1 as the ones-digit answer then carry the 2.

```
   6  7
x  2  3
   ₂  1
```

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

Step 2. Cross-multiply and add: **(6x3)** + **(7x2)** = 32 (tens-digit answer).

use the 2+2carried as the tens-digit answer then carry the 3.

```
   6  7
x  2  3
₃  4  1
```

Trace out the oval ***five times*** with your index finger of your ***non-writing*** hand.

Step 3. Multiply and add: (6x2) + 3carried= 15 (hundreds-digit answer).

```
     6  7
x    2  3
1 5  4  1
```

Answer: 1541

Practice Question 4

86 x 43 = ?

Answer: _________

Self-Reported Page (2 pages)

Cognitive load and motivation measurement questionnaire

All of the following questions refer to the lesson that you just finished. Please respond to each of the questions about the learning activity using the following scale. Please rate the way you felt during the lesson you just studied by ***circling* or ticking** the appropriate number.

not at all applicable								*fully applicable*
0	1	2	3	4	5	6	7	8

1) I enjoyed doing this activity very much

0 1 2 3	4	5 6 7 8

2) This activity was fun to do.

0 1 2 3	4	5 6 7 8

3) I would describe this activity as very interesting.

0 1 2 3	4	5 6 7 8

4) I thought this activity was quite enjoyable.

0 1 2 3	4	5 6 7 8

5) While I was doing this activity, I was thinking about how much I enjoyed it.

0 1 2 3	4	5 6 7 8

6) The learning content was difficult to understand.

0 1 2 3	4	5 6 7 8

7) The explanations of the learning content were difficult to understand.

0 1 2 3	4	5 6 7 8

8) The learning contents were complex.

0 1 2 3	4	5 6 7 8

9) The learning content included much complex information.

0 1 2 3	4	5 6 7 8

10) Without prior knowledge, the information was not understandable.

0 1 2 3	4	5 6 7 8

not at all applicable								fully applicable
0	1	2	3	4	5	6	7	8

11) It was difficult to gain an overview of the structure of the learning material.

0	1	2	3	4	5	6	7	8

12) The design of the learning material made it difficult to recognise links between individual information units.

0	1	2	3	4	5	6	7	8

13) The design of the learning material was inconvenient.

0	1	2	3	4	5	6	7	8

14) The design of the learning material made it difficult to find relevant information quickly.

0	1	2	3	4	5	6	7	8

15) Because of the design of the learning material, I had the impression that I could not concentrate on the learning content.

0	1	2	3	4	5	6	7	8

16) I actively reflected upon the learning content.

0	1	2	3	4	5	6	7	8

17) I consciously focused to understand the learning content.

0	1	2	3	4	5	6	7	8

18) I achieved a comprehensive understanding of the learning content.

0	1	2	3	4	5	6	7	8

19) I was able to expand my prior knowledge with the learning content.

0	1	2	3	4	5	6	7	8

20) I can apply the knowledge that I acquired through the learning material quickly and accurately.

0	1	2	3	4	5	6	7	8

Immediate and Delayed Post-test Page (20 questions in 20 small pages)

1. 56 x 84 = ?	2. 62 x 54 = ?
3. 71 x 73 = ?	4. 75 x 87 = ?

5. 63 x 55 = ?

6. 58 x 64 = ?

7. 52 x 77 = ?

8. 56 x 68 = ?

9. 74 x 65 = ?

10. 82 x 56 = ?

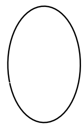
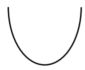
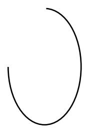
11. 58 x 85 = ?

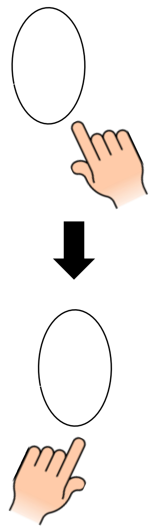
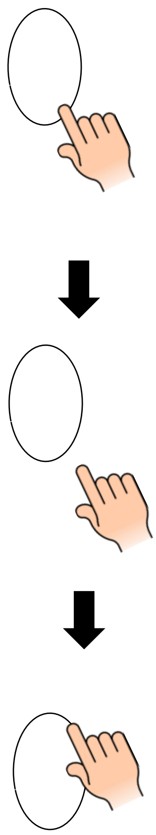
12. 96 x 83 = ?

13. 78 x 93 = ?	14. 76 x 68 = ?
15. 63 x 58 = ?	16. 92 x 87 = ?

17. 86 x 77 = ?	18. 93 x 68 = ?
19. 76 x 96 = ?	20. 82 x 67 = ?

B. Tracing Dosage Video Coding Scheme

Category	Description of Category	Code	Count of Data	Illustration
A	Number of dominant hands' elliptical tracing made on figures as instructed	Dominant hands' elliptical tracing	1 = one tracing action of the ellipse's complete circumference 0.25 = one tracing action of the ellipse's one-quarter circumference 0.5 = one tracing action of the ellipse's two quarters (half an ellipse) circumference 0.75 = one tracing action of the ellipse's three-quarters circumference	1  0.25 
B	Number of non-dominant hands' elliptical tracing actions made on figures as instructed	Non-dominant hands' elliptical tracing	1 = one tracing action of the ellipse's complete circumference 0.25 = one tracing action of the ellipse's one-quarter circumference 0.5 = one tracing action of the ellipse's two quarters (half an ellipse) circumference 0.75 = one tracing action of the ellipse's three-quarters circumference	0.5  0.75 
C	Number of times the participants learned and traced the entire worked example	Learning times	1 = one complete learning process that includes all three steps. 0.33 = one incomplete learning process that includes only one step 0.67 = one incomplete learning process that includes only two steps	

D	Number of elliptical tracing actions of both hands made on figures as instructed	Total hand’s elliptical tracing	D = A + B	
E	Number of times the participants switched between their two hands’ index fingers to trace	Switched times	1 = participants stop the dominant hand’s finger tracing, and then switch to the non-dominant hand’s finger tracing Or 1 = participants stop the non-dominant hand’s finger tracing, and then switch to the dominant hand’s finger tracing	
F	Number of times the participants paused or changed the figures during the same index finger tracing	Paused times	1 = participants use the same index finger to trace Ellipse 1, and then trace Ellipse 2 without fingers changed during it (include participant finished tracing STEP 3 and then learned again from STEP 1) Or 1 = participants stop tracing Ellipse 1 and move the finger to other place, or move the finger away from the paper, and then move the finger back to trace the same Ellipse 1 again. (Not include participant just pause tracing without moving finger away)	

* Any tracing that not on the ellipse or other finger movement like pointing is not

coded

C. Full Set of Experimental Materials (Experiment 2)

Demographic Page (1 page)

Participant No. ___________
Date:

Welcome and General Instruction Phase

This experiment is conducted by Xufei Zhang, Ph.D. student from The University of Sydney.

The experiment is approved by the Human Research Ethics Committee of University of Sydney.

The study is about learning to mental math. The following parts of the study will be presented on the paper. Please read the instruction carefully and complete the tests as required.

Demographic Question Phase

Please answer the following questions

1) What is your age? ____________ years

2) What is your gender? (Please circle one answer.)

A. Female B. Male C. Other (please specify) _____________ D. Prefer not to say

Please choose your preferred hand in the following situations

Please tick one box for each question, indicating whether you prefer to use the left-hand, either-hand, or the right-hand for that task. Only tick the 'either' box if one hand is truly no better than the other. Please answer all questions, and even if you have had little experience in a particular task, try imagining doing that.

		Left	**Either**	**Right**
1	With which hand do you write?			
2	In which hand do you prefer to use a spoon when eating?			
3	In which hand do you prefer to hold a toothbrush when cleaning your teeth?			
4	In which hand do you hold a match when you strike it?			
5	In which hand do you prefer to hold the rubber when erasing a pencil mark?			
6	In which hand do you hold the needle when you are sewing?			
7	When buttering bread, which hand holds the knife?			
8	In which hand do you hold a hammer?			
9	In which hand do you hold the peeler when peeling an apple?			
10	Which hand do you use to draw?			

Pre-test Page (3 pages)

Addition

Part 1 (1 minute)

28 +13	51 +10	42 +53	71 +11	95 +52	74 +38	14 +19	99 +63	57 +83	17 +39
19 +27	40 +44	67 +38	98 +59	42 +13	17 +19	90 +82	45 +91	55 +58	83 +42
98 +31	34 +22	20 +54	63 +92	40 +59	26 +89	18 +39	27 +36	44 +80	88 +77
47 +17	23 +48	41 +53	47 +85	59 +16	23 +18	87 +58	31 +53	38 +49	34 +78
41 +38	86 +93	58 +34	25 +77	86 +55	29 +22	74 +31	34 +19	15 +26	83 +19
37 +98	13 +87	38 +67	51 +65	78 +45	89 +32	34 +65	56 +45	23 +43	47 +39

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

Subtraction

Part 2 (1 minute)

89 -60	52 -48	60 -39	51 -28	85 -23	18 -11	49 -37	83 -57	42 -23	68 -47
52 -19	91 -23	60 -31	42 -31	94 -45	98 -64	50 -33	53 -19	61 -45	41 -27
39 -23	61 -37	29 -19	31 -14	54 -12	92 -65	60 -43	43 -27	70 -31	94 -24
48 -19	42 -31	95 -65	81 -62	40 -31	51 -27	42 -18	97 -18	93 -45	74 -23
65 -39	36 -22	80 -46	51 -27	82 -31	91 -64	68 -59	75 -34	64 -24	42 -37
99 -45	61 -27	91 -60	82 -47	25 -19	73 -45	58 -32	57 -17	59 -42	31 -27

DO NOT GO BACK AND

DO NOT TURN THE PAGE UNTIL ASKED TO DO SO.

Multiplication

Part 3 (1 minute)

73 x 8	41 x 5	69 x 3	29 x 9	16 x 8	63 x 8	60 x 4	52 x 4	85 x 6	36 x 7
52 x 9	98 x 3	41 x 8	19 x 6	15 x 4	49 x 2	71 x 9	30 x 8	48 x 7	81 x 5
45 x 9	32 x 6	79 x 2	37 x 8	19 x 9	52 x 6	17 x 5	47 x 2	39 x 3	78 x 7
14 x 3	38 x 8	50 x 2	80 x 9	61 x 7	52 x 7	97 x 4	72 x 7	16 x 5	49 x 6
52 x 7	71 x 2	96 x 8	47 x 6	83 x 3	16 x 8	44 x 3	50 x 8	62 x 6	39 x 5
13 x 8	68 x 4	75 x 5	67 x 8	45 x 9	94 x 7	52 x 6	83 x 7	61 x 6	54 x 9

DO NOT GO BACK TO ANY OTHER PAGE

Learning Page (Non-Alternating Condition, 10 pages)

Acquisition phase

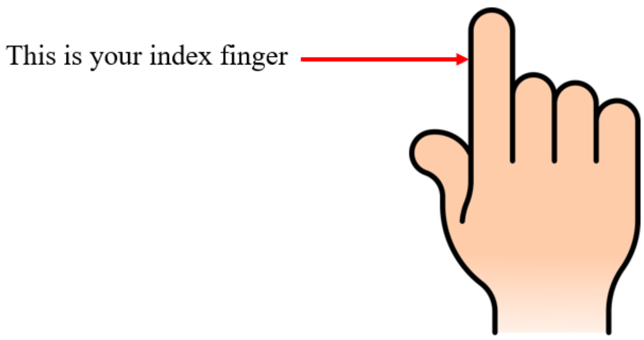
You will now be shown 4 worked examples to study. Please **use your index finger of your writing hand** to help you learn. For each worked example, you will have 2 minutes to:

(1) Look at the worked example;

(2) Read the solution steps in the worked example carefully;

(3)As the instructions in the brackets tell you, use your index finger to trace the operation oval in the worked example on the page.

Please make sure you concentrate on this task because you will be given a very similar problem to solve immediately after each worked example.



Practice: How to add three-digit numbers in a vertical equation

Example: 153 + 164 = ?

$$\begin{array}{r} 153 \\ +\,164 \\ \hline \end{array}$$

Step 1. Add the ones digits: **3 + 4 = 7** (ones-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

$$\begin{array}{r} 1\;5\;3 \\ +\;1\;6\;4 \\ \hline 7 \end{array}$$

Step 2. Add the tens digit: **5 + 6 = 11** (tens-digit answer).
use the 1 as the tens-digit answer then carry the 1.

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

$$\begin{array}{r} 1\;5\;3 \\ +\;1\;6\;4 \\ \hline \boxed{1}\;1\;7 \end{array}$$

Step 3. Add the hundreds digit: **(1+1) + 1carried= 3** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

$$\begin{array}{r} 1\;5\;3 \\ +\;1\;6\;4 \\ \hline 3\;1\;7 \end{array}$$

Answer: =317

Worked Example 1: 12 x 23 = ?

```
  1 2
x 2 3
```

Step 1. Multiply the ones digits: **2 x 3 = 6** (ones-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

```
  1 2
x 2 3
-----
    6
```

Step 2. Cross-multiply and add: **(1x3)** + **(2x2)** = **7** (tens-digit answer).

Trace out the circles each ***five times*** with your index finger of your ***writing*** hand.

```
  1 2
x 2 3
-----
  7 6
```

Step 3. Multiply and add: **1 x 2 = 2** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

```
    1 2
  x 2 3
-------
  2 7 6
```

Answer: 276

Practice Question 1

32 x 21 = ?

Answer: _________

Worked Example 2: 72 x 54 = ?

72
x 54

Step 1. Multiply the ones digits: **2 x 4 = 8** (ones-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

7 2
x 5 4
8

Step 2. Cross-multiply and add: **(7x4)** + **(2x5)** = **38** (tens-digit answer).
Use the **8** as the tens-digit answer then carry the **3**.

Trace out the circles each ***five times*** with your index finger of your ***writing*** hand.

7 2
x 5 4
3 8 8

Step 3. Multiply and add: **(7x5) + 3carried= 38** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

7 2
x 5 4
3 8 8 8

Answer: 3888

Practice Question 2

61 x 47 = ?

Answer: _________

Worked Example 3: 46 x 23 = ?

```
 4 6
x2 3
```

Step 1. Multiply the ones digits: **6 x 3 = 18** (ones-digit answer).

Use the 8 as the ones-digit answer then carry the 1.

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

```
    4  6
x   2  3
-------
   [1] 8
```

Step 2. Cross-multiply and add: **(4x3) + (6x2) = 24** (tens-digit answer).

Use the **4+1carried=5** as the tens-digit answer then carry the **2**.

Trace out the circles each ***five times*** with your index finger of your ***writing*** hand.

```
       4  6
x      2  3
----------
[2]    5  8
```

Step 3. Multiply and add: **(4x2) + 2carried= 10** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

```
       4  6
x      2  3
----------
1  0   5  8
```

Answer: 1058

Practice Question 3

86 x 43 = ?

Answer: _________

Worked Example 4: 78 x 27 = ?

78
x27

Step 1. Multiply the ones digits: **8 x 7 = 56** (ones-digit answer).

Use the 6 as the ones-digit answer then carry the 5.

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

7 8
x 2 7
5 6

Step 2. Cross-multiply and add: **(7x7) + (8x2) = 65** (tens-digit answer).

Use the **5+5carried=(1)0** as the tens-digit answer then carry the **6+(1)=7**

Trace out the circles each ***five times*** with your index finger of your ***writing*** hand.

7 8
x 2 7
7 0 6

Step 3. Multiply and add: **(7x2) + 7carried= 21** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

7 8
x 2 7
2 1 0 6

Answer: 2106

Practice Question 4

56 x 45 = ?

Answer: _________

Learning Page (Alternating Condition, 10 pages)

Acquisition phase

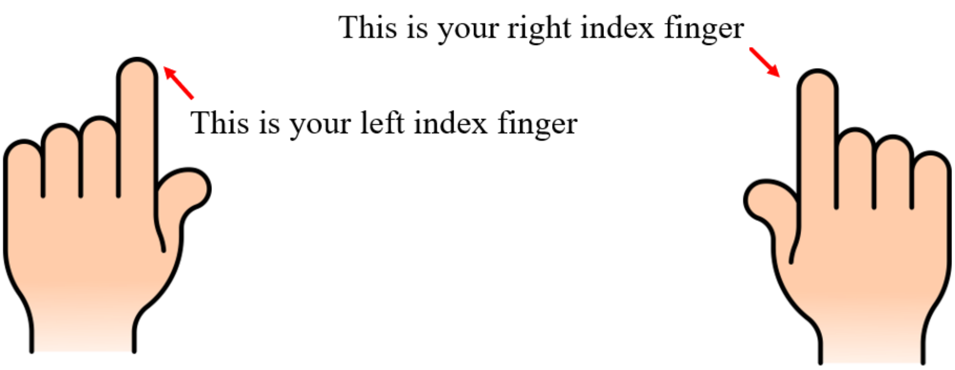
You will now be shown 4 worked examples to study. please use **your finger of your writing or non-writing hand** to help you learn. For each worked example, you will have 2 minutes to:

(1) Look at the worked example;

(2) Read the solution steps in the worked example carefully;

(3) As the instructions in the brackets tell you, use your index finger of your hands to trace the operation oval in the worked example on the page.

Please make sure you concentrate on this task because you will be given a very similar problem to solve immediately after each worked example.



Practice: How to add three-digit numbers in a vertical equation

Example: 153 + 164 = ?

$$\begin{array}{r} 1\,5\,3 \\ +\,1\,6\,4 \\ \hline \end{array}$$

Step 1. Add the ones digits: **3 + 4 = 7** (ones-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

$$\begin{array}{r} 1\;5\;3 \\ +\;1\;6\;4 \\ \hline 7 \end{array}$$

Step 2. Add the tens digit: **5 + 6 = 11** (tens-digit answer).
use the 1 as the tens-digit answer then carry the 1.

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.

$$\begin{array}{r} 1\;5\;3 \\ +\;1\;6\;4 \\ \hline \boxed{1}\;1\;7 \end{array}$$

Step 3. Add the hundreds digit: **(1+1) + 1carried= 3** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

$$\begin{array}{r} 1\;5\;3 \\ +\;1\;6\;4 \\ \hline 3\;1\;7 \end{array}$$

Answer: 317

Worked Example 1: 12 x 23 = ?

```
  1 2
 x2 3
```

Step 1. Multiply the ones digits: **2 x 3 = 6** (ones-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

```
    1  2
 x  2  3
 -------
       6
```

Step 2. Cross-multiply and add: (**1x3**) + (**2x2**) = 7 (tens-digit answer).

Trace out the circles each ***five times*** with your index finger of your ***non-writing*** hand.

```
    1  2
 x  2  3
 -------
    7  6
```

Step 3. Multiply and add: **1 x 2 = 2** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.

```
    1  2
 x  2  3
 -------
 2  7  6
```

Answer: 276

Practice Question 1

32 x 21 = ?

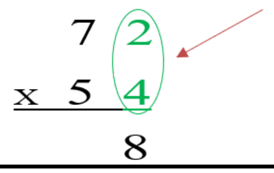
Answer: _________

Worked Example 2: 72 x 54 = ?

```
  72
x 54
```

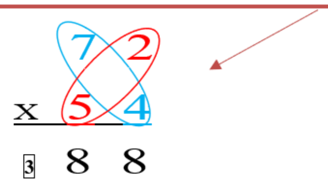
Step 1. Multiply the ones digits: **2 x 4 = 8** (ones-digit answer).

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.



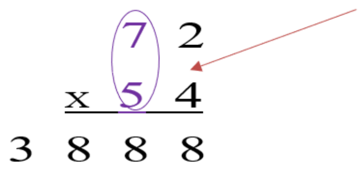
Step 2. Cross-multiply and add: (**7x4**) + (**2x5**) = **38** (tens-digit answer).
Use the **8** as the tens-digit answer then carry the **3**.

Trace out the circles each ***five times*** with your index finger of your ***writing*** hand.



Step 3. Multiply and add: **(7x5) + 3carried= 38** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.



Answer: 3888

Practice Question 2

61 x 47 = ?

Answer: _________

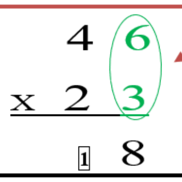
Worked Example 3: 46 x 23 = ?

```
  4 6
x 2 3
```

Step 1. Multiply the ones digits: **6 x 3 = 18** (ones-digit answer).

Use the 8 as the ones-digit answer then carry the 1.

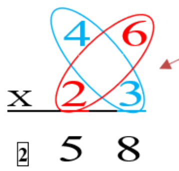
Trace out the circle ***five times*** with your index finger of your ***writing*** hand.



Step 2. Cross-multiply and add: **(4x3)** + **(6x2)** = **24** (tens-digit answer).

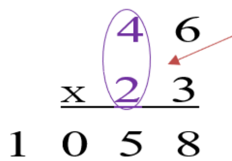
Use the **4+1carried=5** as the tens-digit answer then carry the **2**.

Trace out the circles each ***five times*** with your index finger of your ***non-writing*** hand.



Step 3. Multiply and add: **(4x2) + 2carried= 10** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***writing*** hand.



Answer: 1058

Practice Question 3

86 x 43 = ?

Answer: _________

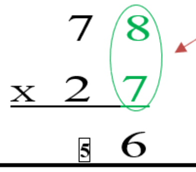
Worked Example 4: 78 x 27 = ?

  78
x 27

Step 1. Multiply the ones digits: **8 x 7 = 56** (ones-digit answer).

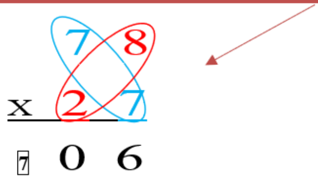
Use the 6 as the ones-digit answer then carry the 5.

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.



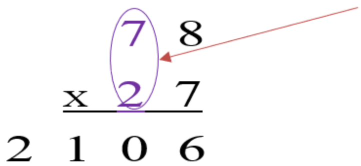
Step 2. Cross-multiply and add: **(7x7)** + **(8x2)** = **65** (tens-digit answer).
Use the **5+5carried=(1)0** as the tens-digit answer then carry the **6+(1)=7**

Trace out the circles each ***five times*** with your index finger of your ***writing*** hand.



Step 3. Multiply and add: **(7x2) + 7carried= 21** (hundreds-digit answer).

Trace out the circle ***five times*** with your index finger of your ***non-writing*** hand.



Answer: 2106

Practice Question 4

56 x 45 = ?

Answer: _________

Self-Reported Page (2 pages)

Cognitive load and motivation measurement questionnaire

All of the following questions refer to the lesson that you just finished. Please respond to each of the questions about the learning activity using the following scale. Please rate the way you felt during the lesson you just studied by ***circling* or ticking** the appropriate number.

not at all applicable								*fully applicable*
0	1	2	3	4	5	6	7	8

1) I enjoyed doing this activity very much

0	1	2	3	4	5	6	7	8

2) This activity was fun to do.

0	1	2	3	4	5	6	7	8

3) I would describe this activity as very interesting.

0	1	2	3	4	5	6	7	8

4) I thought this activity was quite enjoyable.

0	1	2	3	4	5	6	7	8

5) While I was doing this activity, I was thinking about how much I enjoyed it.

0	1	2	3	4	5	6	7	8

6) The learning content was difficult to understand.

0	1	2	3	4	5	6	7	8

7) The explanations of the learning content were difficult to understand.

0	1	2	3	4	5	6	7	8

8) The learning contents were complex.

0	1	2	3	4	5	6	7	8

9) The learning content included much complex information.

0	1	2	3	4	5	6	7	8

10) Without prior knowledge, the information was not understandable.

0	1	2	3	4	5	6	7	8

`

not at all applicable								*fully applicable*
0	1	2	3	4	5	6	7	8

11) It was difficult to gain an overview of the structure of the learning material.

0	1	2	3	4	5	6	7	8

12) The design of the learning material made it difficult to recognise links between individual information units.

0	1	2	3	4	5	6	7	8

13) The design of the learning material was inconvenient.

0	1	2	3	4	5	6	7	8

14) The design of the learning material made it difficult to find relevant information quickly.

0	1	2	3	4	5	6	7	8

15) Because of the design of the learning material, I had the impression that I could not concentrate on the learning content.

0	1	2	3	4	5	6	7	8

16) I actively reflected upon the learning content.

0	1	2	3	4	5	6	7	8

17) I consciously focused to understand the learning content.

0	1	2	3	4	5	6	7	8

18) I achieved a comprehensive understanding of the learning content.

0	1	2	3	4	5	6	7	8

19) I was able to expand my prior knowledge with the learning content.

0	1	2	3	4	5	6	7	8

20) I can apply the knowledge that I acquired through the learning material quickly and accurately.

0	1	2	3	4	5	6	7	8

Immediate and Delayed Post-test Page (20 questions in 20 small pages)

1) 56 x 74 = ?	2) 62 x 54 = ?
3) 71 x 73 = ?	4) 75 x 87 = ?

5) 63 x 55 = ?	6) 58 x 62 = ?
7) 52 x 73 = ?	8) 56 x 68 = ?

9) 74 x 62 = ?	10) 82 x 56 = ?
11) 58 x 85 = ?	12) 96 x 83 = ?

13) 78 x 93 = ?	14) 76 x 68 = ?
15) 63 x 53 = ?	16) 92 x 87 = ?

17) 86 x 77 = ?	18) 91 x 68 = ?
19) 76 x 96 = ?	20) 82 x 67 = ?

D. Full Set of Experimental Materials (Experiment 3)

Demographic Page (1 page)

Participant No. ___________
Date:

Welcome and General Instruction Phase

This experiment is conducted by Xufei Zhang, Ph.D. student from The University of Sydney.

The experiment is approved by the Human Research Ethics Committee of University of Sydney.

The study is about learning to mental math. The following parts of the study will be presented on the paper. Please read the instruction carefully and complete the tests as required.

Demographic Question Phase

Please answer the following questions

1) What is your age? ____________ years

2) What is your gender? (Please circle one answer.)

A. Female B. Male C. Other (please specify) _____________ D. Prefer not to say

Please choose your preferred hand in the following situations

Please tick one box for each question, indicating whether you prefer to use the left-hand, either-hand, or the right-hand for that task. Only tick the 'either' box if one hand is truly no better than the other. Please answer all questions, and even if you have had little experience in a particular task, try imagining doing that.

		Left	**Either**	**Right**
1	With which hand do you write?			
2	In which hand do you prefer to use a spoon when eating?			
3	In which hand do you prefer to hold a toothbrush when cleaning your teeth?			
4	In which hand do you hold a match when you strike it?			
5	In which hand do you prefer to hold the rubber when erasing a pencil mark?			
6	In which hand do you hold the needle when you are sewing?			
7	When buttering bread, which hand holds the knife?			
8	In which hand do you hold a hammer?			
9	In which hand do you hold the peeler when peeling an apple?			
10	Which hand do you use to draw?			

Pre-test Page (3 pages)

Addition

Part 1 (1 minute)

28 +13	51 +10	42 +53	71 +11	95 +52	74 +38	14 +19	99 +63	57 +83	17 +39
19 +27	40 +44	67 +38	98 +59	42 +13	17 +19	90 +82	45 +91	55 +58	83 +42
98 +31	34 +22	20 +54	63 +92	40 +59	26 +89	18 +39	27 +36	44 +80	88 +77
47 +17	23 +48	41 +53	47 +85	59 +16	23 +18	87 +58	31 +53	38 +49	34 +78
41 +38	86 +93	58 +34	25 +77	86 +55	29 +22	74 +31	34 +19	15 +26	83 +19
37 +98	13 +87	38 +67	51 +65	78 +45	89 +32	34 +65	56 +45	23 +43	47 +39

DO NOT TURN THIS PAGE UNTIL ASKED TO DO SO.

Subtraction

Part 2 (1 minute)

89 -60	52 -48	60 -39	51 -28	85 -23	18 -11	49 -37	83 -57	42 -23	68 -47
52 -19	91 -23	60 -31	42 -31	94 -45	98 -64	50 -33	53 -19	61 -45	41 -27
39 -23	61 -37	29 -19	31 -14	54 -12	92 -65	60 -43	43 -27	70 -31	94 -24
48 -19	42 -31	95 -65	81 -62	40 -31	51 -27	42 -18	97 -18	93 -45	74 -23
65 -39	36 -22	80 -46	51 -27	82 -31	91 -64	68 -59	75 -34	64 -24	42 -37
99 -45	61 -27	91 -60	82 -47	25 -19	73 -45	58 -32	57 -17	59 -42	31 -27

DO NOT GO BACK AND

DO NOT TURN THE PAGE UNTIL ASKED TO DO SO.

Multiplication

Part 3 (1 minute)

73 x 8	41 x 5	69 x 3	29 x 9	16 x 8	63 x 8	60 x 4	52 x 4	85 x 6	36 x 7
52 x 9	98 x 3	41 x 8	19 x 6	15 x 4	49 x 2	71 x 9	30 x 8	48 x 7	81 x 5
45 x 9	32 x 6	79 x 2	37 x 8	19 x 9	52 x 6	17 x 5	47 x 2	39 x 3	78 x 7
14 x 3	38 x 8	50 x 2	80 x 9	61 x 7	52 x 7	97 x 4	72 x 7	16 x 5	49 x 6
52 x 7	71 x 2	96 x 8	47 x 6	83 x 3	16 x 8	44 x 3	50 x 8	62 x 6	39 x 5
13 x 8	68 x 4	75 x 5	67 x 8	45 x 9	94 x 7	52 x 6	83 x 7	61 x 6	54 x 9

DO NOT GO BACK TO ANY OTHER PAGE

Learning and Self-Report Page (Non-Alternating Condition, 18 pages)

Acquisition phase

You will now be shown 6 worked examples to study. All 6 worked examples will teach you the same mental math method. Please **use your index finger of your right hand** to help you learn. For each worked example, you will have 2 minutes to:

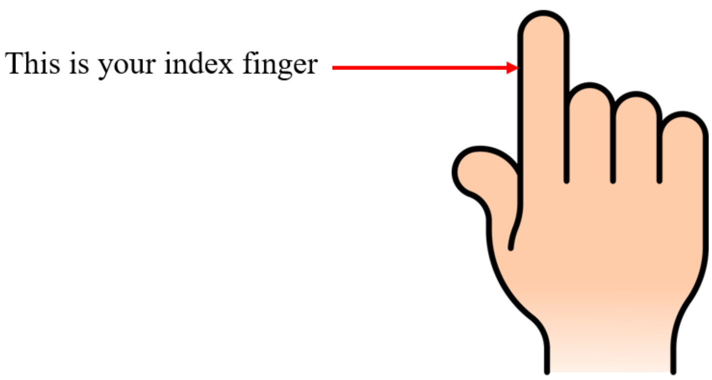
(1) Look at the worked example;

(2) Read the solution steps in the worked example carefully;

(3) Follow picture's instruction, use your index finger to trace the operation oval in the worked example.

You need to solve a very similar problem immediately after each worked example.

Please make sure you concentrate during the whole 2 minutes.



Tracing Practice: How to add three-digit numbers in a vertical equation

Example: 153 + 164 = ?

```
  1 5 3
+ 1 6 4
```

Step 1. Add the ones digits: **3** + **4** = **7 (ones-digit answer)**.

```
  1 5 3
+ 1 6 4
-------
      7
```

Right Hand.
Five Times.

Step 2. Add the tens digit: **5** + **6** = **11 (tens-digit answer)**.
use the 1 as the tens-digit answer then carry the 1.

```
  1 5 3
+ 1 6 4
-------
  1 1 7
```

Right Hand.
Five Times.

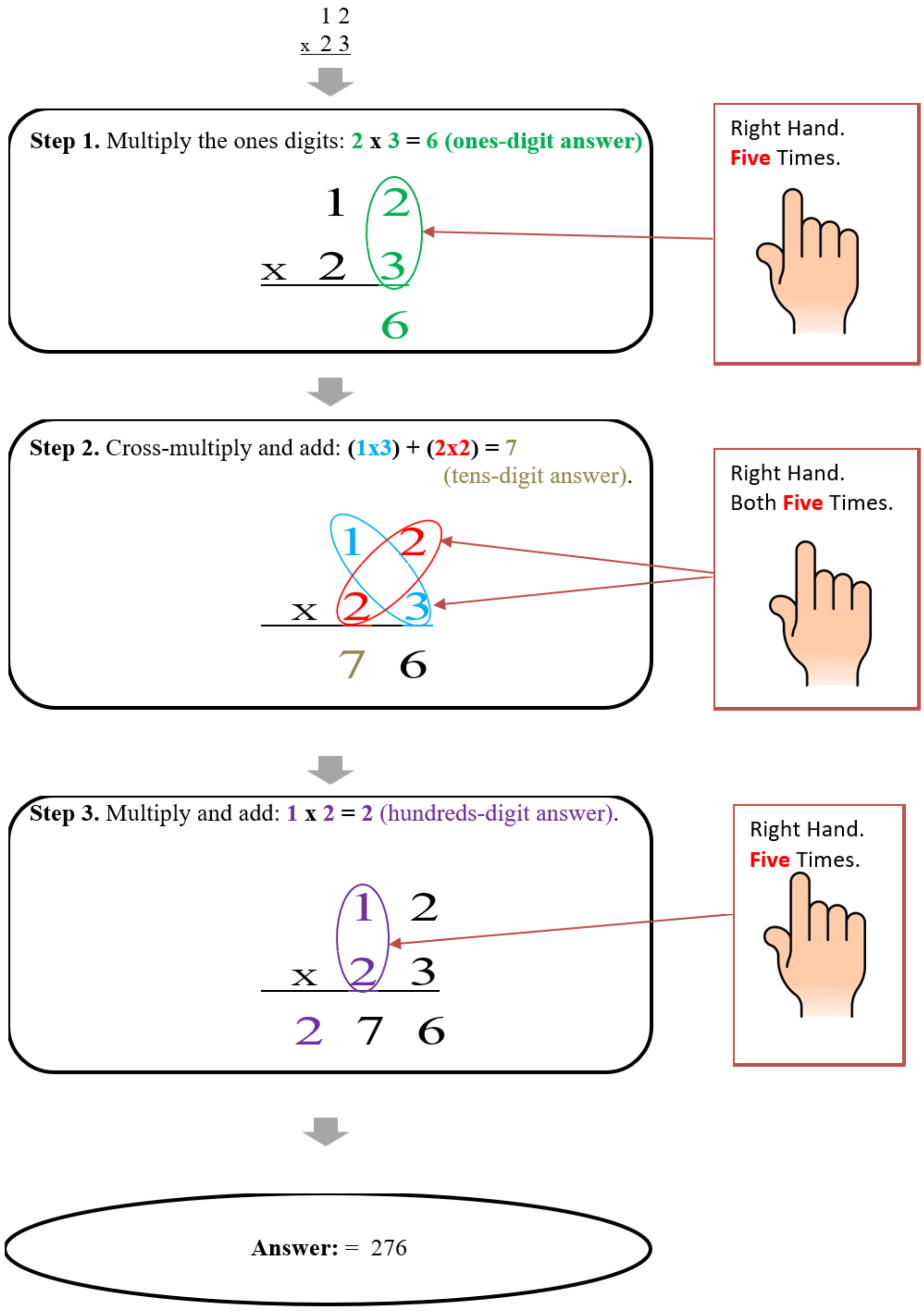
Step 3. Add the hundreds digit: (**1**+**1**) + **1 carried** = **3 (hundreds-digit answer)**

```
  1 5 3
+ 1 6 4
-------
  3 1 7
```

Right Hand.
Five Times.

Answer: =317

Worked Example 1: 12 x 23 = ?

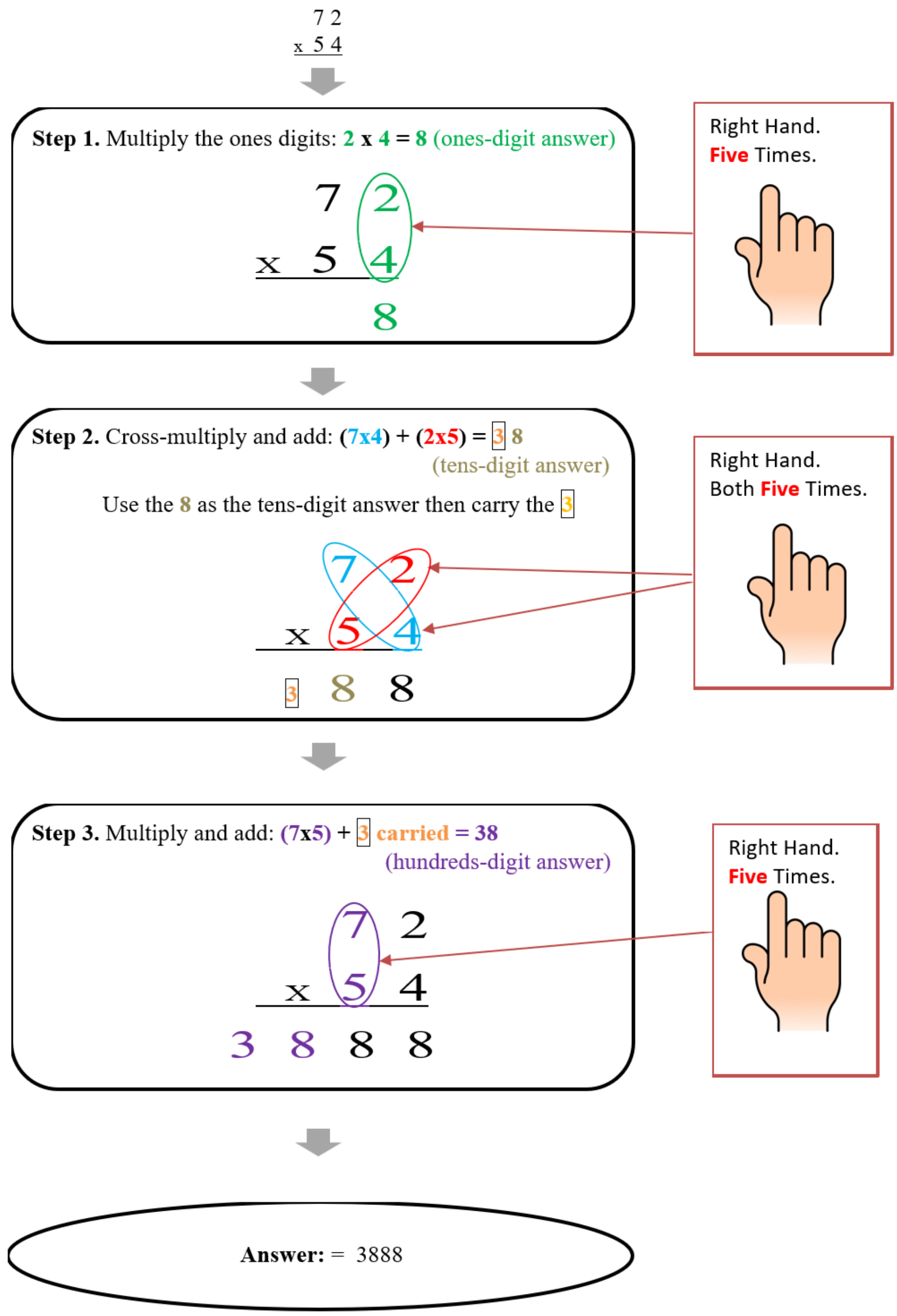


Practice Question 1

14 x 21 = ?

Answer: __________

Worked Example 2: 72 x 54 = ?

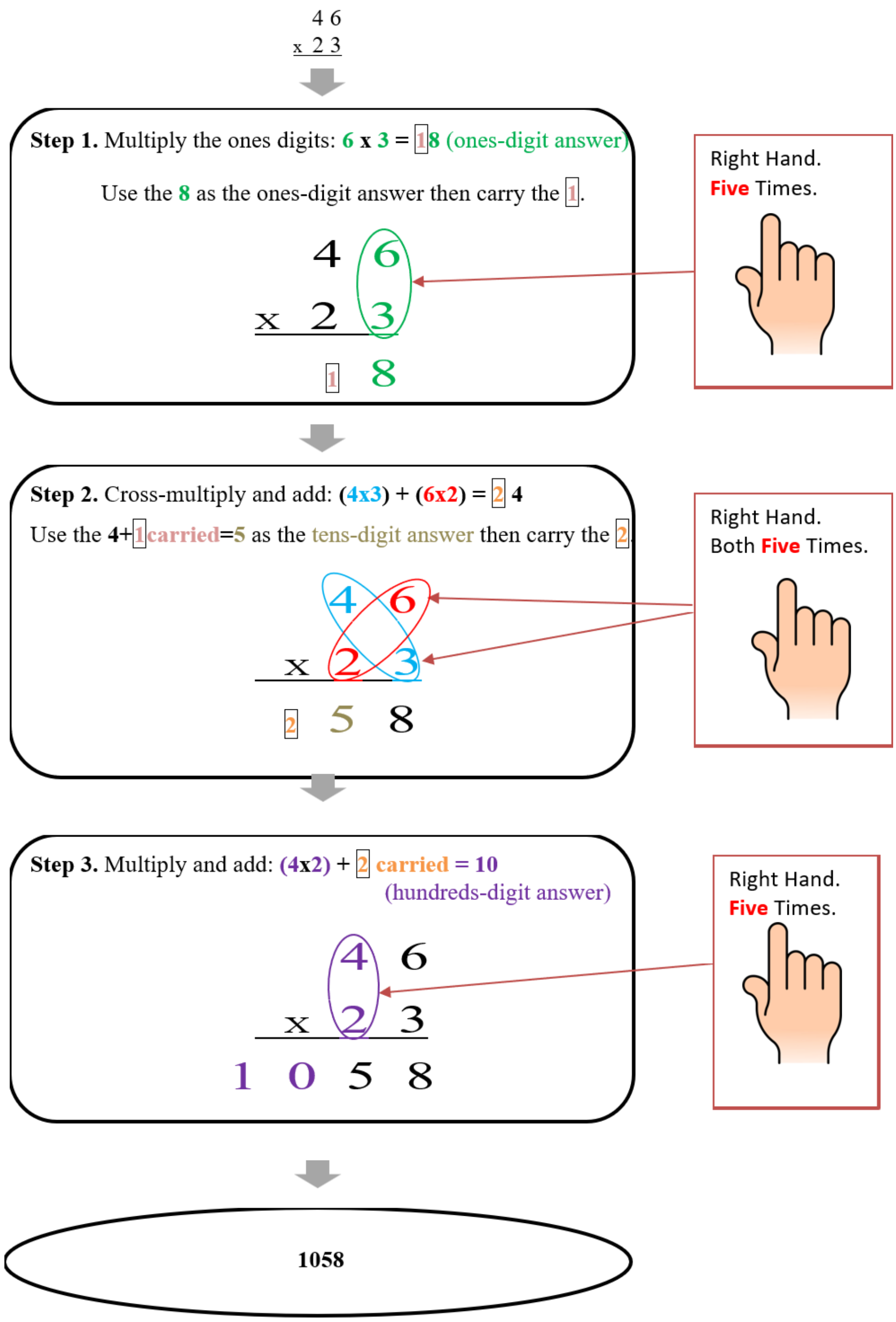


Practice Question 2

61 x 47 = ?

Answer: _________

Worked Example 3: 46 x 23 = ?



Practice Question 3

86 x 43 = ?

Answer: _________

Cognitive load and motivation measurement questionnaire

Please respond to the following questions according to the three worked examples (worked example 1-3) that you just finished.

1) I enjoyed doing this activity very much

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

2) This activity was fun to do.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

3) I would describe this activity as very interesting.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

4) I thought this activity was quite enjoyable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

5) While I was doing this activity, I was thinking about how much I enjoyed it.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

6) The learning content was difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

7) The explanations of the learning content were difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

8) The learning contents were complex.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

9) The learning content included much complex information.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

10) Without prior knowledge, the information was not understandable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

11) It was difficult to gain an overview of the structure of the learning material.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

12) The design of the learning material made it difficult to recognise links between individual information units.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

13) The design of the learning material was inconvenient.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

14) The design of the learning material made it difficult to find relevant information quickly.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

15) Because of the design of the learning material, I had the impression that I could not concentrate on the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

16) I actively reflected upon the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

17) I consciously focused to understand the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

18) I achieved a comprehensive understanding of the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

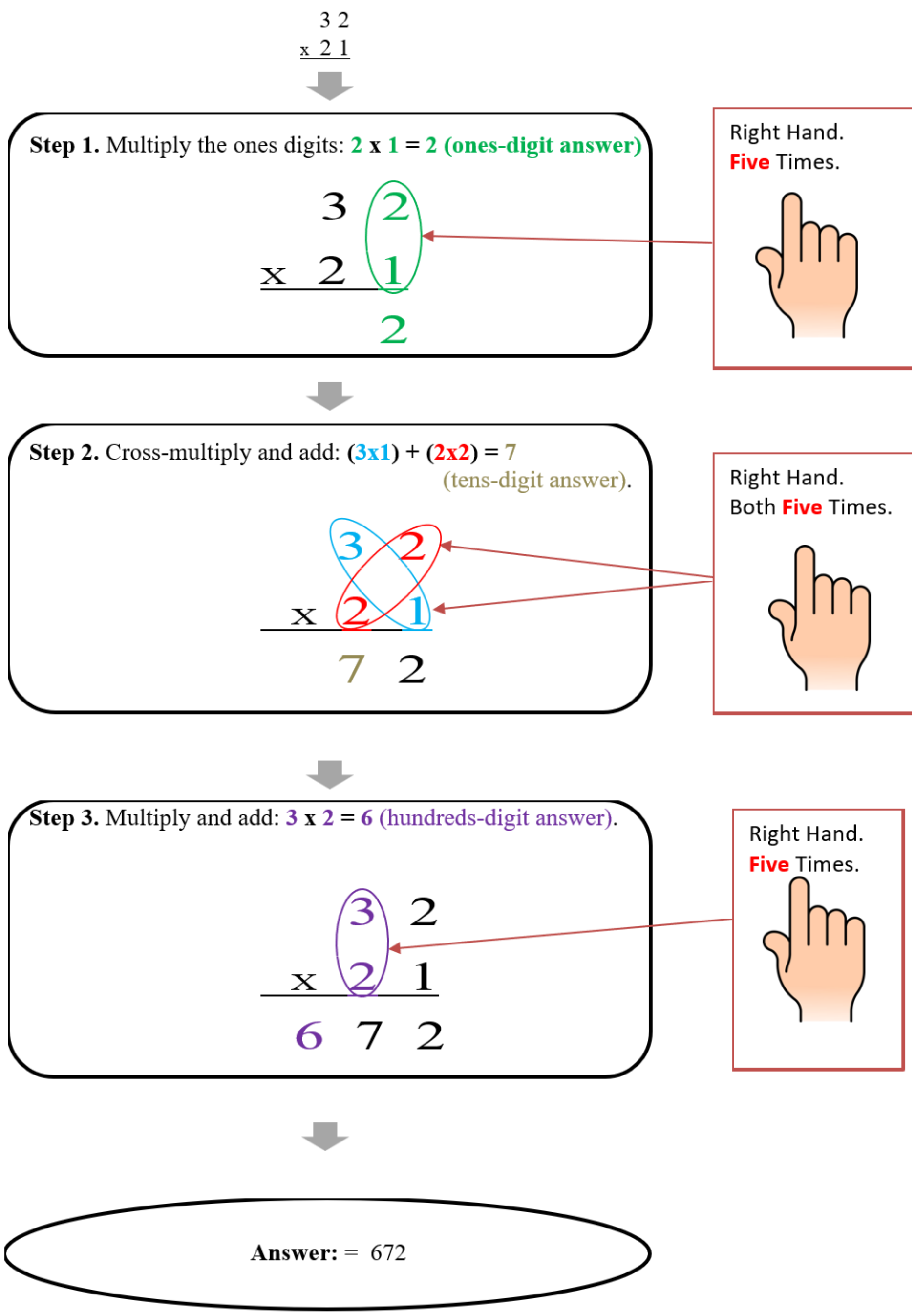
19) I was able to expand my prior knowledge with the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

20) I can apply the knowledge that I acquired through the learning material quickly and accurately.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

Worked Example 4: 32 x 21 = ?



Practice Question 4

33 x 12 = ?

Answer: __________

Worked Example 5: 52 x 63 = ?

```
  5 2
x 6 3
```

Step 1. Multiply the ones digits: **2 x 3 = 6** (ones-digit answer)

```
  5 2
x 6 3
    6
```

Step 2. Cross-multiply and add: (**5x3**) + (**2x6**) = 2 7
(tens-digit answer)

Use the 7 as the tens-digit answer then carry the 2

```
  5 2
x 6 3
2 7 6
```

Step 3. Multiply and add: **(5x6)** + 2 **carried = 32**
(hundreds-digit answer)

```
    5 2
  x 6 3
3 2 7 6
```

Answer: = 3276

Practice Question 5

54 x 32 = ?

Answer: ___________

Worked Example 6: 35 x 64 = ?

```
  3 5
x 6 4
```

Step 1. Multiply the ones digits: **5 x 4 = 20** (ones-digit answer)

Use the **0** as the ones-digit answer then carry the 2.

```
    3  5
x   6  4
    2  0
```

Right Hand.
Five Times.

Step 2. Cross-multiply and add: **(3x4) + (5x6) = 42**

Use the **2+2carried=4** as the tens-digit answer then carry the 4.

```
    3  5
x   6  4
 4  4  0
```

Right Hand.
Both **Five** Times.

Step 3. Multiply and add: **(3x6) + 4 carried = 22** (hundreds-digit answer)

```
       3  5
x      6  4
 2  2  4  0
```

Right Hand.
Five Times.

2240

Practice Question 6

65 x 82 = ?

Answer: _________

Cognitive load and motivation measurement questionnaire

Please respond to the following questions according to the three worked examples (worked example 4-6) that you just finished.

1) I enjoyed doing this activity very much

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

2) This activity was fun to do.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

3) I would describe this activity as very interesting.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

4) I thought this activity was quite enjoyable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

5) While I was doing this activity, I was thinking about how much I enjoyed it.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

6) The learning content was difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

7) The explanations of the learning content were difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

8) The learning contents were complex.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

9) The learning content included much complex information.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

10) Without prior knowledge, the information was not understandable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

11) It was difficult to gain an overview of the structure of the learning material.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

12) The design of the learning material made it difficult to recognise links between individual information units.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

13) The design of the learning material was inconvenient.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

14) The design of the learning material made it difficult to find relevant information quickly.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

15) Because of the design of the learning material, I had the impression that I could not concentrate on the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

16) I actively reflected upon the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

17) I consciously focused to understand the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

18) I achieved a comprehensive understanding of the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

19) I was able to expand my prior knowledge with the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

20) I can apply the knowledge that I acquired through the learning material quickly and accurately.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

Learning and Self-Report Page (Alternating Condition, 18 pages)

Acquisition phase

You will now be shown 4 worked examples to study. All 4 worked examples will teach you the same mental math method. Please **use your index finger of your right or left hand** to help you learn. For each worked example, you will have 2 minutes to:

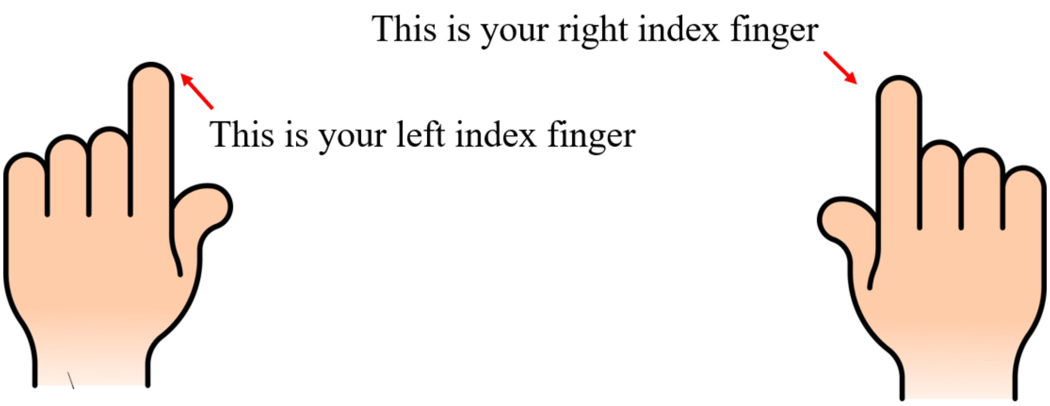
(1) Look at the worked example;

(2) Read the solution steps in the worked example carefully;

(3) Follow picture's instruction, use your index finger to trace the operation oval in the worked example.

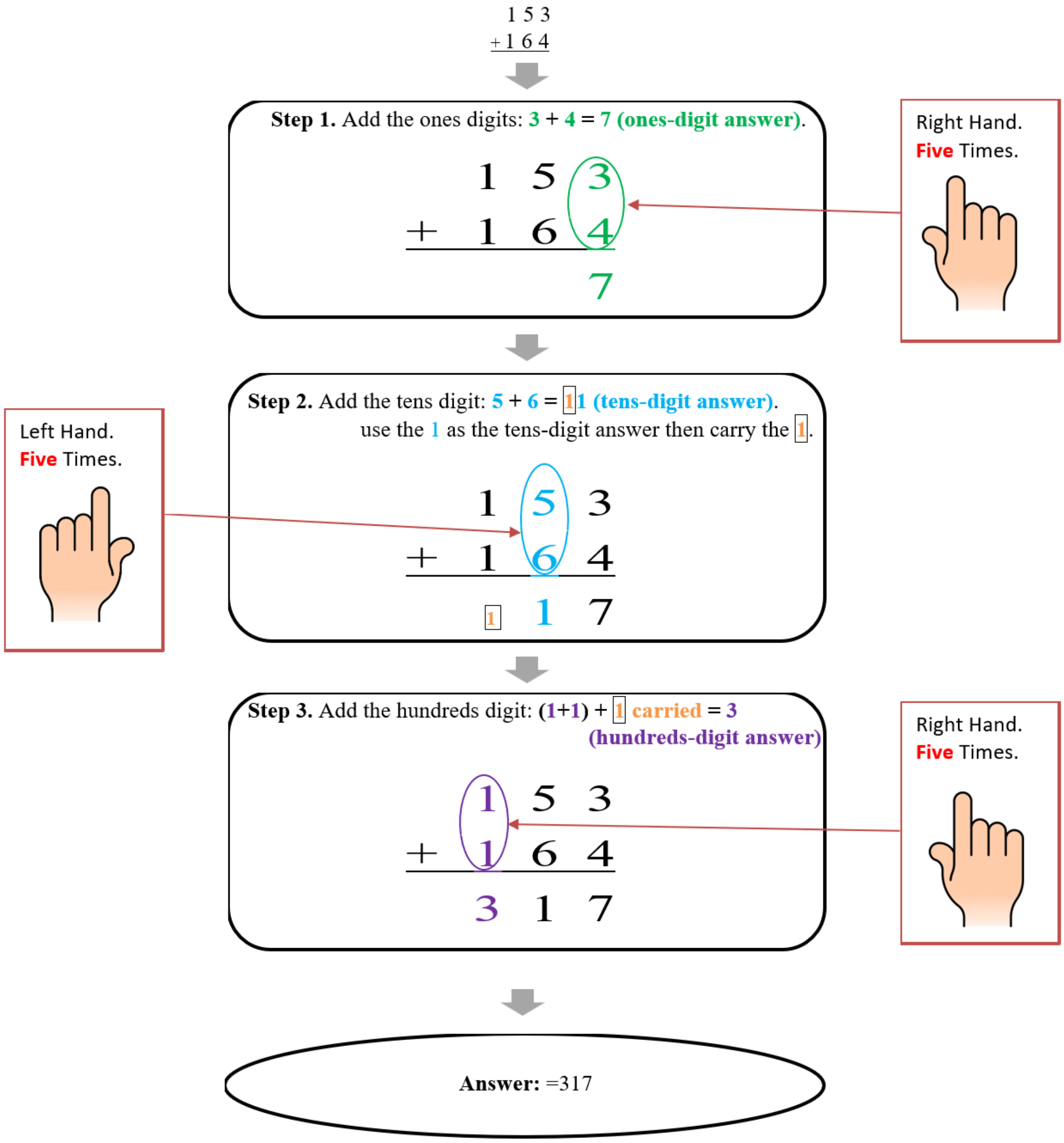
You need to solve a very similar problem immediately after each worked example.

Please make sure you concentrate during the whole 2 minutes.

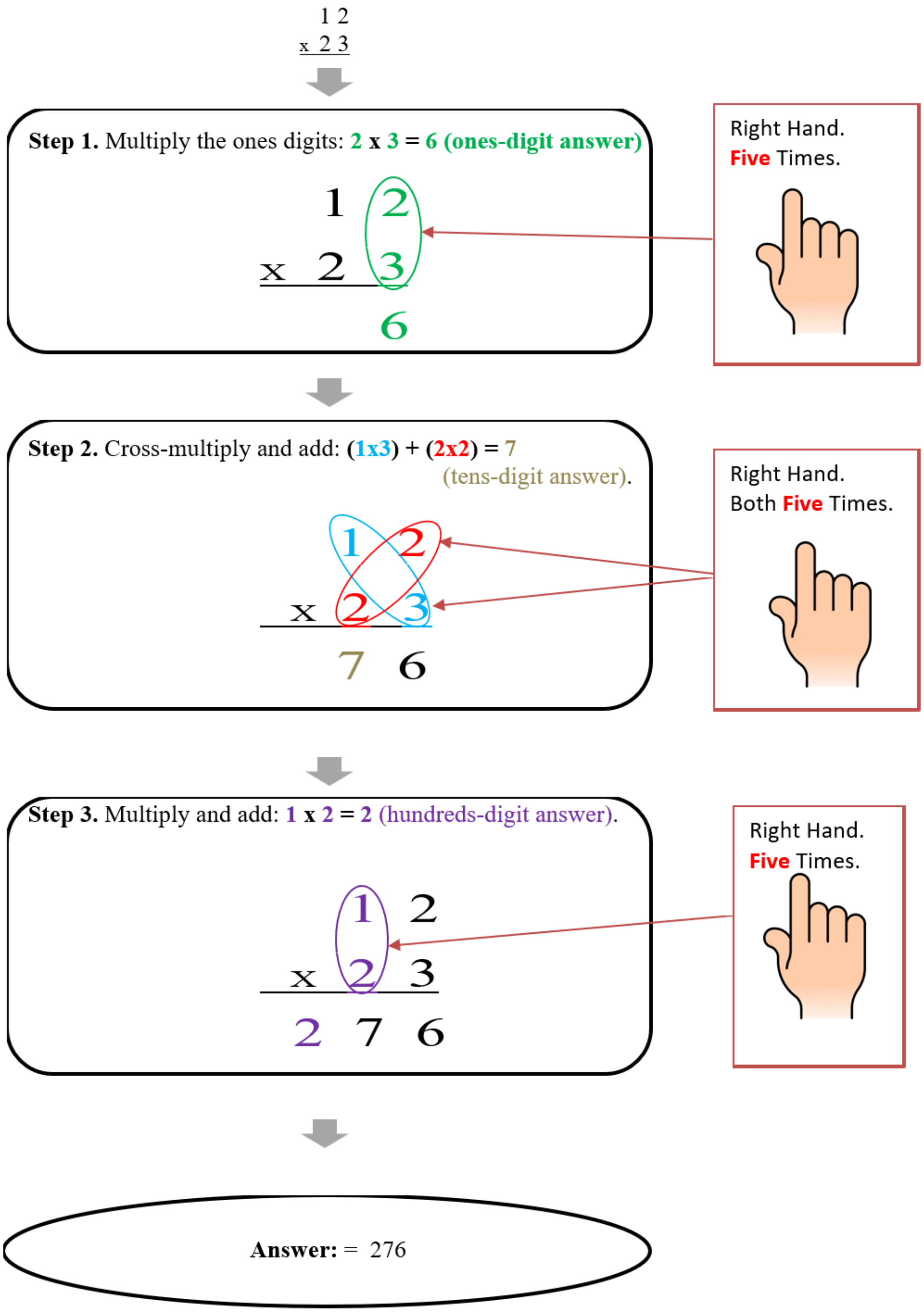


Tracing Practice: How to add three-digit numbers in a vertical equation

Example: 153 + 164 = ?



Worked Example 1: 12 x 23 = ?



Practice Question 1

14 x 21 = ?

Answer: __________

Worked Example 2: 72 x 54 = ?

```
   7 2
 x 5 4
```

Step 1. Multiply the ones digits: **2 x 4 = 8** (ones-digit answer)

```
    7  2
x   5  4
       8
```

Step 2. Cross-multiply and add: (**7x4**) + (**2x5**) = **3 8**
(tens-digit answer)

Use the **8** as the tens-digit answer then carry the 3

```
     7  2
x    5  4
3    8  8
```

Step 3. Multiply and add: (**7x5**) + 3 **carried = 38**
(hundreds-digit answer)

```
       7  2
x      5  4
3   8  8  8
```

Answer: = 3888

Practice Question 2

61 x 47 = ?

Answer: _________

Worked Example 3: 46 x 23 = ?

```
   4 6
 x 2 3
```

Step 1. Multiply the ones digits: **6 x 3** = 1 8 (ones-digit answer)

Use the **8** as the ones-digit answer then carry the 1.

```
     4  6
 x   2  3
     1  8
```

Right Hand.
Five Times.

Step 2. Cross-multiply and add: **(4x3)** + **(6x2)** = 2 **4**

Use the **4**+1 **carried**=**5** as the tens-digit answer then carry the 2.

```
      4  6
 x    2  3
   2  5  8
```

Right Hand.
Both **Five** Times.

Step 3. Multiply and add: **(4x2)** + 2 **carried = 10** (hundreds-digit answer)

```
        4  6
 x      2  3
  1  0  5  8
```

Right Hand.
Five Times.

1058

Practice Question 3

86 x 43 = ?

Answer: _________

Cognitive load and motivation measurement questionnaire

Please respond to the following questions according to the three worked examples (worked example 1-3) that you just finished.

1) I enjoyed doing this activity very much

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

2) This activity was fun to do.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

3) I would describe this activity as very interesting.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

4) I thought this activity was quite enjoyable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

5) While I was doing this activity, I was thinking about how much I enjoyed it.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

6) The learning content was difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

7) The explanations of the learning content were difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

8) The learning contents were complex.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

9) The learning content included much complex information.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

10) Without prior knowledge, the information was not understandable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

11) It was difficult to gain an overview of the structure of the learning material.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

12) The design of the learning material made it difficult to recognise links between individual information units.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

13) The design of the learning material was inconvenient.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

14) The design of the learning material made it difficult to find relevant information quickly.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

15) Because of the design of the learning material, I had the impression that I could not concentrate on the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

16) I actively reflected upon the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

17) I consciously focused to understand the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

18) I achieved a comprehensive understanding of the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

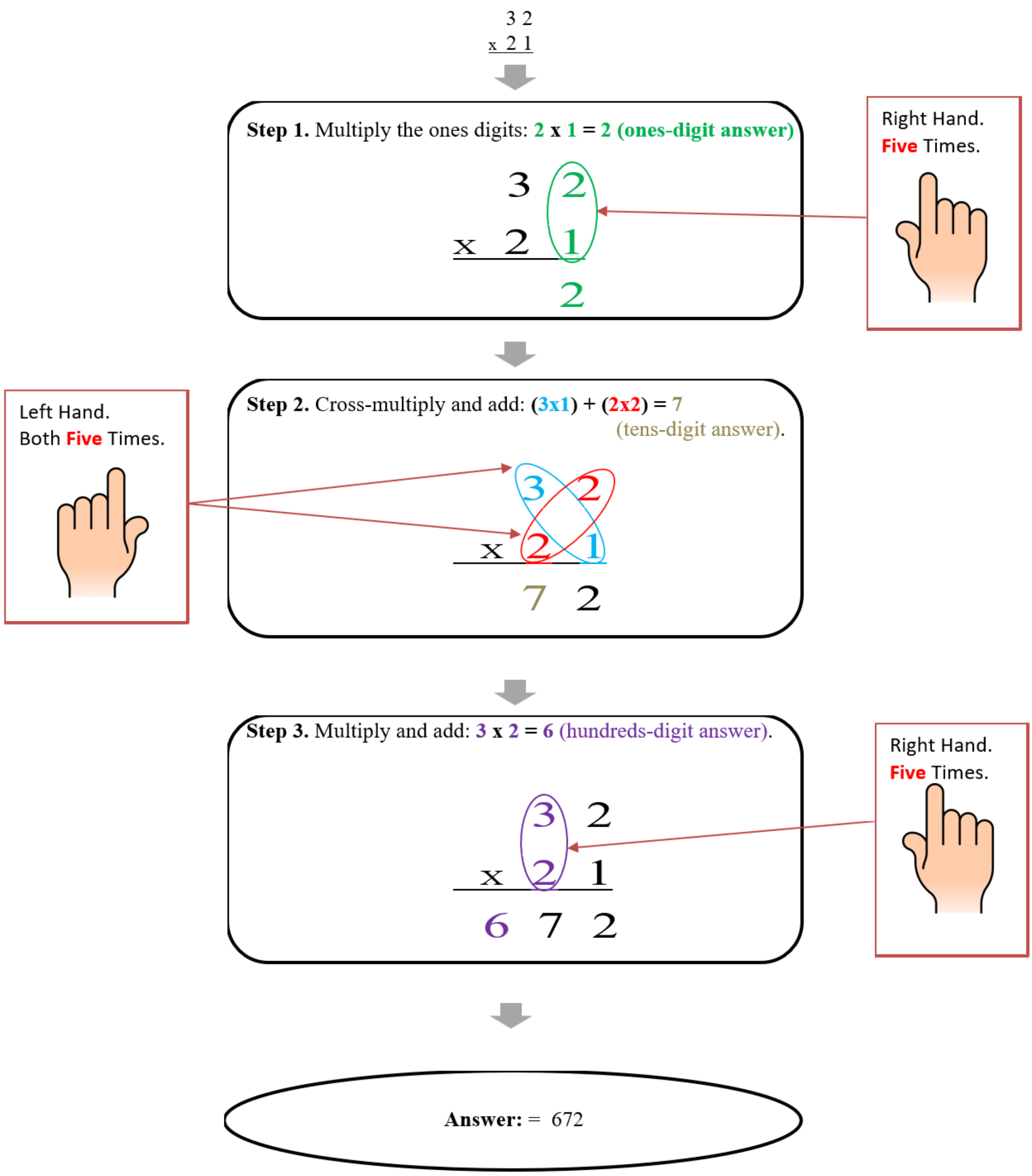
19) I was able to expand my prior knowledge with the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

20) I can apply the knowledge that I acquired through the learning material quickly and accurately.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

Worked Example 4: 32 x 21 = ?



Practice Question 4

33 x 12 = ?

Answer: __________

Worked Example 5: 52 x 63 = ?

```
  5 2
x 6 3
```

Left Hand.
Five Times.

Step 1. Multiply the ones digits: **2 x 3 = 6** (ones-digit answer)

```
  5 2
x 6 3
    6
```

Step 2. Cross-multiply and add: **(5x3)** + **(2x6)** = 2 7
(tens-digit answer)

Use the 7 as the tens-digit answer then carry the 2

```
    5 2
  x 6 3
2   7 6
```

Right Hand.
Both **Five** Times.

Left Hand.
Five Times.

Step 3. Multiply and add: **(5x6)** + 2 **carried = 32**
(hundreds-digit answer)

```
      5 2
    x 6 3
3 2 7 6
```

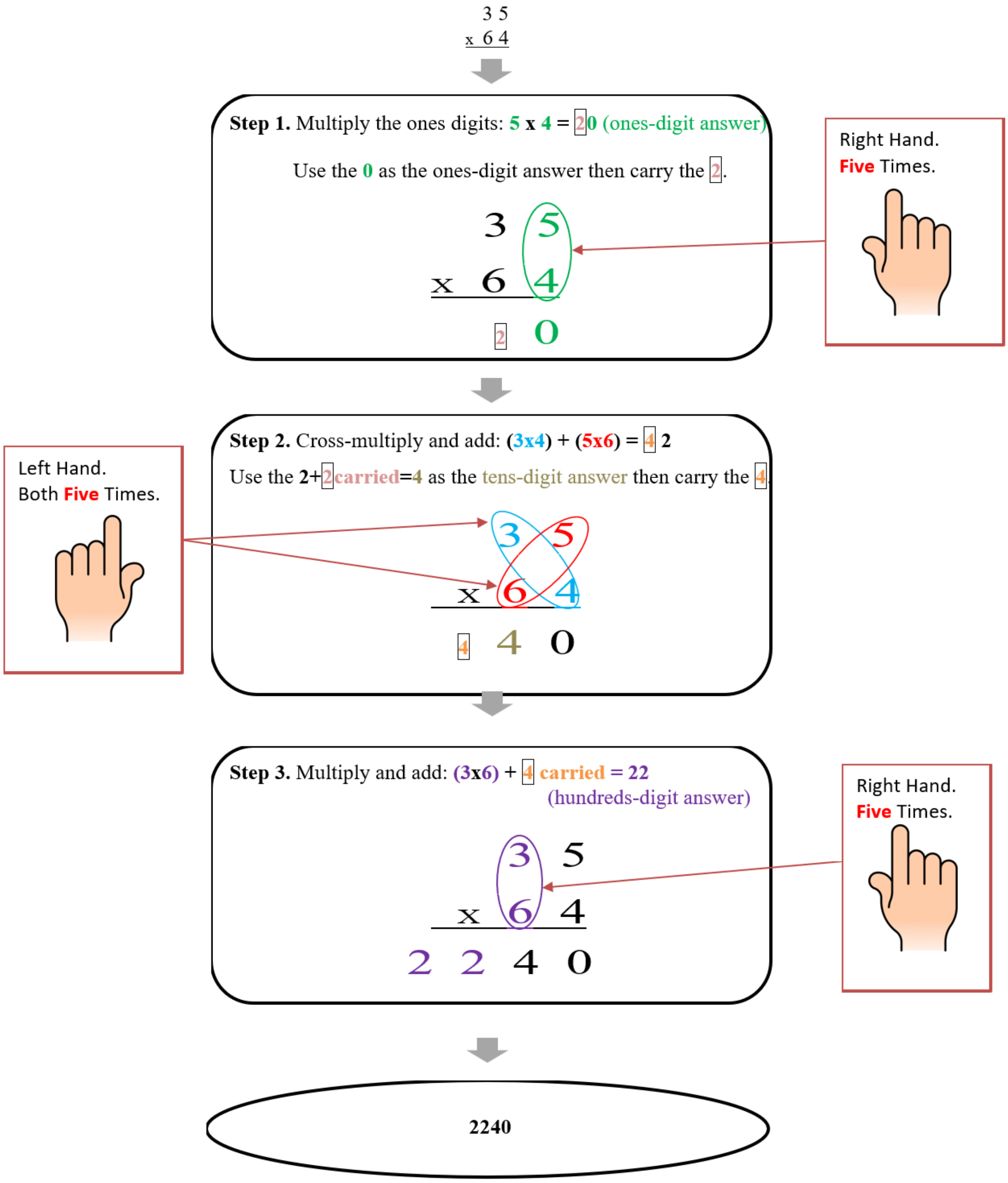
Answer: = 3276

Practice Question 5

54 x 32 = ?

Answer: ___________

Worked Example 6: 35 x 64 = ?



Practice Question 6

65 x 82 = ?

Answer: _________

Cognitive load and motivation measurement questionnaire

Please respond to the following questions according to the three worked examples (worked example 4-6) that you just finished.

1) I enjoyed doing this activity very much

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

2) This activity was fun to do.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

3) I would describe this activity as very interesting.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

4) I thought this activity was quite enjoyable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

5) While I was doing this activity, I was thinking about how much I enjoyed it.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

6) The learning content was difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

7) The explanations of the learning content were difficult to understand.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

8) The learning contents were complex.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

9) The learning content included much complex information.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

10) Without prior knowledge, the information was not understandable.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

11) It was difficult to gain an overview of the structure of the learning material.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

12) The design of the learning material made it difficult to recognise links between individual information units.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

13) The design of the learning material was inconvenient.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

14) The design of the learning material made it difficult to find relevant information quickly.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

15) Because of the design of the learning material, I had the impression that I could not concentrate on the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

16) I actively reflected upon the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

17) I consciously focused to understand the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

18) I achieved a comprehensive understanding of the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

19) I was able to expand my prior knowledge with the learning content.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

20) I can apply the knowledge that I acquired through the learning material quickly and accurately.

not at all applicable	0	1	2	3	4	5	6	7	8	*fully applicable*

Immediate Post-test Page (20 questions in 20 small pages)

1) 56 x 74 = ?	2) 16 x 54 = ?
3) 33 x 74 = ?	4) 75 x 87 = ?

5) 63 x 55 = ?	6) 58 x 62 = ?
7) 54 x 15 = ?	8) 56 x 68 = ?

9) 36 x 75 = ?	10) 82 x 56 = ?
11) 58 x 85 = ?	12) 96 x 83 = ?

13) 78 x 93 = ?	14) 76 x 68 = ?
15) 56 x 77 = ?	16) 92 x 87 = ?

17) 86 x 76 = ?	18) 91 x 68 = ?
19) 76 x 83 = ?	20) 82 x 67 = ?

Delayed Post-test Page (20 questions in 20 small pages)

21) 56 x 63 = ?	22) 76 x 84 = ?
23) 47 x 62 = ?	24) 47 x 56 = ?

25) 56 x 53 = ?	26) 58 x 65 = ?
27) 56 x 95 = ?	28) 75 x 64 = ?

29) 58 x 45 = ?

30) 54 x 73 = ?

31) 67 x 57 = ?

32) 43 x 65 = ?

33) 78 x 23 = ?	34) 67 x 58 = ?
35) 54 x 76 = ?	36) 76 x 95 = ?

37) 74 x 69 = ?

38) 65 x 79 = ?

39) 78 x 63 = ?

40) 76 x 46 = ?